Latent class models for categorical data with a multilevel structure

Lukociene, O.

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Latent class models for categorical data with a multilevel structure

Olga Lukočienė

Ph.D. Disseration Tilburg University

October 2009

Contents

Chapter 1

Introduction

Although the majority of the statistical methods assume independence of observations, social and behavioral science research designs often yield observations that cannot be assumed to be independent. Examples are data from longitudinal studies, data from studies based on multilevel designs, and survey data obtained via cluster-sampling designs (Snijders and Bosker, 1999). In each of these cases, the dependencies arise because of the nested or hierarchical structure of the observations. Dependencies can, however, also occur in the form of non-nested structures, such as in social network and spatial data.

The most popular approach for analyzing dependent observations is by statistical models with random effects. Such models are referred to as hierarchical models, mixed models, multilevel models, and random-coefficient models (Bryk and Raudenbush, 1992; Goldstein, 1995; Hox, 2002; Skrondal and Rabe-Hesketh, 2004; Snijders and Bosker, 1999). Linear regression models with random effects have become a standard statistical tool, especially because general statistical packages (SPSS and

SAS) include routines for this purpose. With specialized packages, one can include random effects in more complicated models, for example, in factor analytic models. Estimation of random-effects models is quite straightforward as long as both prediction errors and random effects can be assumed to come from normal distributions.

However, the two basic assumptions underlying the standard random-effects approaches – normal errors and normal random effects - may be unrealistic in social science research. Outcome variables of interest are very often not continuous but categorical variables, which makes assuming normal errors unrealistic. Variants of random-effects models for categorical dependent variables have been developed, such as random-effects logistic and Poisson regression models (Hartzel *et al.*, 2001; Hedeker and Gibbons, 1996; Rabe-Hesketh *et al.*, 2001; Wong and Mason, 1985). Estimation of these so-called nonlinear mixed models is, however, much more difficult and time consuming than the estimation of linear mixed models. The main issue is that complicated integrals have to be solved, which can be done by numerical integration (Bock and Aitkin, 1981), simulated-likelihood, quasi-likelihood (Goldstein, 1995), or Bayesian simulation (Fox and Glas, 2001) methods. Some of these methods become computationally intensive with more than a few random effects while the others may perform poorly (Agresti *et al.*, 2000; Lesaffre and Spiessens, 2001; Rodriguez and Goldman, 1995, 2001).

The other basic assumption that may not hold in practice is that the random effects come from a multivariate normal distribution. An alternative is to work with a finite mixture distribution. This yields a

so-called nonparametric approach, in which each individual is assumed to belong to one of a small number of latent classes that differ with respect to the parameters of the model of interest (Heckman and Singer, 1982; Laird, 1978; Vermunt, 1997). When applied in the context of regression analysis, one obtains what is usually referred to as latent class regression or mixture regression model (Wedel and DeSarbo, 1994). As pointed out by Aitkin (1999), an advantage of such a nonparametric approach is that it is not necessary to introduce possibly inappropriate and unverifiable assumptions about the distribution of the random effects. Another advantage is that it is much more practical when the regression model is nonlinear: models with several random coefficients that take hours of computation time with a parametric approach are estimated within a minute when adopting a nonparametric approach (Vermunt and Dijk, 2001).

Latent class regression and random coefficients modeling have always been seen as very different approaches for dealing with dependent observations. Recently, Aitkin (1999), Hartzel *et al.* (2001), Rabe-Hesketh *et al.* (2001), Skrondal and Rabe-Hesketh (2004), and Vermunt and Dijk (2001) stressed the connecting between the two approaches and showed that latent class regression methods cannot only be used to identify latent classes with different regression coefficients, but may also yield the standard random-coefficient modeling output; that is, estimates for the fixed and random effects.

A limitation of the latent class regression model has always been that it could only be applied with data sets consisting of no more than two

levels of nesting. Recently, Vermunt (2003, 2004, 2008) showed how to overcome this limitation. This means that the nonparametric random-effects approach can now also be used with more than two levels of nesting. With non-nested dependency structures, things become much more complicated. In such situations, one has to rely on (Bayesian) simulation methods for parameter estimation. This is, however, not specific for the nonparametric approach, but applies to any type of non-hierarchical mixed model with non-normal errors.

Another problem has been the lack of user-friendly software. While software for linear mixed modelling has been generally available for quite some time, availability of software for latent class regression modelling is a very recent development. Several easy to use, as well as slightly more advanced packages are now available, such as GLIMMIX (Wedel, 2001), Latent GOLD (Vermunt and Magidson, 2005), GLLAMM (Rabe-Hesketh *et al.*, 2001), and Mplus (Múthen and Múthen, 2006). This means that the method is generally available for applied researchers, both in academics and business.

1.1 Purpose of this study

One of the purposes of this study is to provide a systematic comparison of the two random-effects approaches. As indicated above, the nonparametric approach has several practical advantages, but which is of course not enough to prefer that particular method. A topic that deserves special attention when using LC-based random-effects models is the decision

about the number of classes. In the context of data sets containing more than two levels of nesting there is an additional complicating factor, namely that it requires the simultaneous decision about the number of classes at multiple levels.

The main questions that will be addressed in this research project are:

1. Is the nonparametric model more reliable in situations in which the assumptions underlying the parametric model do not hold?

2. Does it harm if we use a nonparametric model say for practical reasons when the assumptions of the parametric model hold?

3. How should we determine the number of latent classes in models for two-level and three-level data sets?

Some work has been done related to these questions. Aitkin (1999) compared parametric and nonparametric models for a few small data sets from the biomedical field. A similar thing should be done for social science applications. Andrews *et al.* (2002*a*) and Andrews *et al.* (2002*b*) compared the parametric and nonparametric approaches in a simulation study on conjoint experiments. They, however, concentrated on a somewhat different type of research question; that is, on prediction of the dependent variable instead of recovery of the parameters associated with the fixed and random effects. Their work provides a good example on how to set up our study. Hartzel *et al.* (2001) did a small simulation study for a multinomial logistic model with a random intercept.

Their main conclusion was that more research is needed to provide a final conclusion about the relative performance of the two methods.

The above-mentioned studies indicate that the nonparametric approach performs at least as good as the parametric approach, but that it also has its problematic aspects, two of which are the issue of deciding about the number of latent classes and the occurrence of boundary estimates. This means that in this project special attention should be paid to these complications for which possible solutions have been proposed in the literature.

As far as determining the number classes is concerned, we can on the one hand make a distinction between the nonparametric maximum likelihood (NMPL) estimation procedure, in which this is not an issue because the number of classes is increased till a saturation point, and the latent class regression approach, in which the number of latent classes is typically selected using information criteria (Andrews and Currim, 2003; Dias, 2004). On the other hand, when expanding to more than two levels, additional complications arise such as that multiple dependent decisions are involved and that it is not clear what sample size definition should be used when using information criteria which have the sample size in their formulae.

The best manner to answer our three research questions defined above is by means of simulation studies in which data sets are generated from known populations. It is then investigated whether estimated parameters are close to the population values, after taking sampling fluctuation into account. Data sets are simulated from populations that are typical

for social and behavioural science research, where factors such as distribution of the random effects, sample size, number and size of random effects, and number of fixed effects are varied.

1.2 Outline of this thesis

This dissertation is a collection of four self-contained manuscripts, with as the common topic the use of latent class analysis as a tool for multilevel analysis with categorical responses.

The first two chapters compare the performance of latent class based nonparametric random effects modeling with standard parametric modeling in the context of multilevel binary logistic regression analysis. The evaluation criteria are bias and relative efficiency. Chapter 2 provides a detailed comparison of these two approaches when only the intercept is assumed to be a random effect. Chapter 3 extends Chapter 2 to the more complex situation of multidimensional instead of unidimensional random effects; that is, when not only the intercept but also the slopes are random. In addition, it includes a semi-parametric latent-class based random effects approach in the comparison.

The next two chapters deal with the selection of the number of latent classes in multilevel latent class analysis, which is an example of a three-level model for categorical data. Chapter 4 presents results of a simulation study focusing on the selection of the number classes at the highest hierarchical level under simplifying assumption that the number of lower-level latent classes is known. Chapter 5 extends the work re-

ported in Chapter 4 to the simultaneous decision about the number of higher- and lower-level latent classes. Moreover, it proposes a new three-step model fitting strategy for multilevel latent class analysis, which provides a way to deal with the dependency between the decisions about the number of lower- and higher-level classes.

All four chapters have either been published or submitted for publication, with Olga Lukočienė as the first author and Jeroen K. Vermunt as a co-author. Roberta Varriale serves as another co-author of Chapter 5.

Chapter 2

A comparison of multilevel logistic regression models with parametric and nonparametric random intercepts

2.1 Introduction

In the biomedical, social and behavioral sciences, it is common to collect data with a nested, multilevel, or hierarchical structure. It is therefore not surprising that the last decades there has been an increase in the use of multilevel models in these fields (Hox, 2002; Skrondal and Rabe-Hesketh, 2004; Snijders and Bosker, 1999). Examples of nested data structures include persons nested within families, pupils nested within schools, patients nested within primary care physicians, and repeated measurements nested within subjects. In more general terms, lower-level or level-1 observational units (persons, pupils, patients, or repeated measurements) are nested within higher-level or level-2 observational

units (families, schools, primary care physicians, or subjects).

Specific for multilevel data sets is that observations are correlated; that is, level-1 units (pupils, time points) belonging to the same level-2 unit (schools, subjects) tend to be more alike than level-1 units from different level-2 units. Methods for dealing with correlated data are usually classified as marginal or conditional models (Lee and Nelder, 2004). In marginal models such as the GEE approach by Zeger *et al.* (1988), the correlation between observations is treated as a nuisance factor. In contrast, in conditional models, specification of the dependence structure is part of the model building. Random effects models – sometimes also referred to as subject-specific models – belong to the family of conditional models, since results are conditional on the level-2 units' unobserved random effects. A limitation of random effects models that may be problematic in particular types of applications is that these can only capture positive associations between nested observations. Alternative conditional models which can also yield negative associations are, for example, transition models in which a person's state at a particular time point is modeled conditional on the state at the previous time point.

In this research, we focus on conditional models which use random effects. Whereas initially random effects were introduced for linear regression models, currently they were also applied in combination with the more general class of generalized linear models, yielding what is often referred to as the family of generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993; Stiratelli *et al.*, 1984) or hierarchical gen-

eralized linear models (HGLMs) (Lee and Nelder, 2004). Usually the unobserved random effects are assumed to come from a particular parametric distribution, typically multivariate normal (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993). But it is clear that parametric distributional assumptions about the random effects are unlikely to hold in practice (Aitkin, 1999). Various studies found that misspecification of the distribution of random effects results in a light loss of efficiency of the regression estimators (Heagerty and Kurland, 2001; Maas and Hox, 2004; Neuhaus *et al.*, 1992).

As an alternative to using a mixing distribution from a parametric family, one may use a nonparametric specification for the random effects distribution (Heckman and Singer, 1982; Laird, 1978). This involves using a discrete mixing distribution defined by a set of unknown locations and weights to approximate an underlying continuous mixing distribution with an unknown form. Maximum likelihood (ML) estimation of the resulting finite mixture or latent class model is straightforward using the Expectation-Maximization (EM) algorithm: since the likelihood is a finite mixture no (numerical) integration is involved. By choosing the number of latent classes to maximize the likelihood, the nonparametric maximum likelihood (NPML) estimator is obtained (Böhning, 2000; Heckman and Singer, 1984; Laird, 1978). When used in the context of regression analysis, one obtains what is sometimes referred to as a latent class or mixture regression model (Leisch, 2004; Vermunt and Dijk, 2001; Wedel and DeSarbo, 1994).

Latent class and random coefficients regression models have always

been seen as rather different approaches for dealing with dependent observations. Recently, the connection between the two approaches was stressed and it was shown that latent class regression methods cannot only be used to identify latent classes with different regression coefficients, but may also yield the typical random-coefficient modelling output; that is, estimates for the fixed and random effects (Aitkin, 1999; Hartzel *et al.*, 2001; Rabe-Hesketh *et al.*, 2005; Vermunt and Dijk, 2001). As pointed out by Aitkin (1999), an important advantage of such a nonparametric approach is that there is no need to introduce possibly inappropriate and unverifiable assumptions about the distribution of the random effects. But this is certainly not enough to prefer this particular method, which is generally available in mixture modelling software such as GLLAMM (Skrondal and Rabe-Hesketh, 2004), Latent GOLD (Vermunt and Magidson, 2005), and Mplus (Múthen and Múthen, 1998).

Based on a limited scope simulation study for a random intercept ordinal logit model, Hartzel *et al.* (2001) concluded tentatively that the parametric approach yields more reliable estimates for both the fixed and random intercept terms, although it had some difficulties when the random effects distribution was extremely skewed. For the remaining fixed effects both approaches produce essentially unbiased estimates. They indicated that more research is needed to provide a final conclusion about the relative performance of the two methods. In contrast, based on another small simulation study for three types of GLMMs, Agresti *et al.* (2004) advised always to use a nonparametric instead of a parametric specification for the random effects distribution in order to prevent loss

of efficiency. Though the simulation studies by Hartzel *et al.* (2001) and Agresti *et al.* (2004) seem to yield contradictory conclusions, closer inspection of their designs provides a possible explanation for the encountered differences. Hartzel *et al.* (2001) used small lower-level sample sizes (4 and 7) combined with a moderate higher-level sample size (100) and small values of the random effects variances. Agresti *et al.* (2004) used moderate to large lower-level sample sizes (10, 20, and 100) combined with small higher-level sample sizes (10 and 30) and moderate to large random effects variances. Our hypothesis is that the differences in conclusions are the result of these differences in simulation set up, and that lower-level and higher-level sample sizes and random effects variances should be more systematically varied to provide a complete answer.

This paper provides such a more systematic comparison of the two random effects approaches for the two-level random intercept logistic regression model. More specifically, the two research questions that are addressed are:

1. Should the nonparametric model be preferred in situations in which underlying assumptions of the parametric model do not hold?

2. Does it harm using a nonparametric model – say for practical reasons – when the assumptions of the parametric model hold?

A simulation study was conducted in which a broad range of data sets were generated in order to cover all typical populations in biomedical, social, and behavioral science research. More specifically, we varied the true distribution of the random effects, the size of the intraclass correla-

tion coefficient (ICC), and the level-1 and level-2 sample sizes. We are interested in whether these simulation design factors affect the answers to our two research questions.

The next section describes the models of interest. Section 2.3 discusses the set up of the simulation study. Results of the simulation study are presented in Section 2.4. In Section 2.5, we present an application of the parametric and nonparametric random effects logistic regression model to a real life data set. The last section provides the reader with a discussion along with conclusions and practical recommendations.

2.2 The two-level random-intercept model

This section introduces two-level generalized linear models with either a parametric or a nonparametric random intercept. Let y_{ij} denote the observed response of the level-1 unit i, $i = 1, \ldots, n_j$, belonging to level-2 unit j, $j = 1, \ldots, n$, $\boldsymbol{x}_{ij}$ the vector of explanatory variables, and u_j the unobservable common random effect for all level-1 units within level-2 unit j. The vector $\boldsymbol{x}_{ij}$ may contain different types of explanatory variables; that is, variables that vary between level-1 units, between level-2 units, or between both level-1 and level-2 units, as well as (cross-level) interaction terms. In a GLMM, the conditional mean of y_{ij}, $E[y_{ij}|\boldsymbol{x}_{ij}, u_j]$, denoted by μ_{ij} is related to the linear predictor as follows:

$$g(\mu_{ij}) = \boldsymbol{\beta}'\boldsymbol{x}_{ij} + u_j, \tag{2.1}$$

where $g(\cdot)$ is what is referred to as the link function. Note that this is the special case in which only the intercept is random.

The typical specification for the random intercept term u_j, $j = 1, \ldots, n$, is to assume that this is an independently and identically distributed normal random variable with mean zero and variance σ_u^2; that is, $u_j \sim N(0, \sigma_u^2)$. An equivalent alternative is to treat the mean as a free parameter and fix the β for the intercept to 0. Consistent with this distributional assumption, parameters of GLMMs are usually estimated by ML, where construction of the likelihood function is simplified by the fact that y_{ij} can be assumed to be independent within level-2 units conditionally on the observed predictors and the unobserved random effects. ML estimation involves maximizing the following marginal likelihood function:

$$L(\boldsymbol{\beta}, \sigma_u^2) = \prod_{j=1}^{n} \int_u \left[\prod_{i=1}^{n_j} f(y_{ij}|\boldsymbol{x}_{ij}, u; \boldsymbol{\beta}) \right] f(u; \sigma_u^2) du, \tag{2.2}$$

where $f(y_{ij}|\boldsymbol{x}_{ij}, u; \boldsymbol{\beta})$ represents the error distribution at level-1 or, equivalently, the conditional density of y_{ij}. Note that the fixed effects $\boldsymbol{\beta}$ and the variance σ_u^2 are the unknown parameters to be estimated. Except for the situation in which a continuous response variable is modelled with an identity link function and a normal level-1 error distribution, maximization of the likelihood requires the optimization of a numerically integrated likelihood. For this numerical integration, one may use a technique called Gauss-Hermite quadrature, which uses an optimal discrete approximations of the normal distribution. The most common algorithms for maximizing the resulting numerically integrated

marginal likelihood are the EM algorithm (Agresti *et al.*, 2000; Bock and Aitkin, 1981; Dempster *et al.*, 1977) and gradient methods, such as the Fisher scoring (Longford, 1987) and Newton-Raphson algorithm (Pan and Thompson, 2003; Rabe-Hesketh *et al.*, 2004). In our study we used numerical integration with 50 nodes. For maximization a combination of EM and Newton-Raphson was used, where the estimation process starts with EM iterations and switches to Newton-Raphson when the relative change in the parameter values is very small.

As was indicated in the introduction, usually nothing or very little it is known about the underlying distribution of the random effects. To prevent possible misspecification, it may therefore be attractive to assume the random effects u_j to come from an unspecified mixing distribution concentrated on a finite number of latent classes or mass points (Aitkin, 1999; Böhning, 2000; Heckman and Singer, 1984; Laird, 1978). Let K denote the number of latent classes, k a particular latent class, and u_k^* the unknown value of the random effect u_j when level-2 unit j belongs to latent class k, and let $\pi_k = P(u_j = u_k^*)$ represent the probability that a randomly selected level-2 unit belongs to latent class k. Using such a K-class discrete characterization of the random effects distribution yields the following marginal likelihood function:

$$L(\boldsymbol{\beta}, \boldsymbol{u^*}, \boldsymbol{\pi}) = \prod_{j=1}^{n} \sum_{k=1}^{K} \prod_{i=1}^{n_j} f(y_{ij}|\boldsymbol{x}_{ij}, u_j = u_k^*; \boldsymbol{\beta})\, \pi_k, \tag{2.3}$$

where $f(y_{ij}|\boldsymbol{x}_{ij}, u_j = u_k^*; \boldsymbol{\beta})$ is the class-specific conditional density function of y_{ij}. Note that $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$, and that moreover one

identifying location constraint should be imposed on the u_k^* parameters, e.g., $\sum_{k=1}^{K} u_k^* \pi_k = 0$, which implies that the u_k^* are centered. The unknown parameters to be estimated are the fixed effects $\boldsymbol{\beta}$, $K-1$ free mass point locations u_k^*, and $K-1$ free mass point weights π_k. Even though the random effects variance itself is not a model parameter, it can easily be obtained as follows: $\sigma_{u^*}^2 = \sum_{k=1}^{K} (u_k^*)^2 \pi_k$.

Maximization of the marginal likelihood function in equation (2.3) for a specific K can, as in the parametric case, be achieved by means of the EM and/or Newton-Raphson algorithm. The use of multiple sets of starting values is usually required because of the risk of ending up in a local maximum.

In a standard finite mixture modelling context one estimates the model of interest for different values of K and stops increasing the number of classes when the model fit no longer improves according to the BIC, AIC or another criterion. However, to obtain the solution corresponding to the NPML estimate of the random effects distribution, we not only have to maximize (2.3) for specific values of K, but we simultaneously have to find the value of K – say K_{NPML} – that yields the largest marginal likelihood value. In other words, we have to find the saturation point at which increasing K no longer results in an increase of the likelihood function. A method to find K_{NPML} proposed by various authors involves introducing latent classes one by one using directional (Gateaux) derivatives (Böhning, 2000; Lindsay, 1983, 1995; Rabe-Hesketh *et al.*, 2003). A much simpler alternative approach is to estimate the model with a large number of latent classes, K_{MAX}. When

$K_{MAX} > K_{NPML}$, the ML estimates for u_k^* will be equal for some latent classes and/or the estimate for π_k will be equal to zero for some latent classes (Böhning, 2000). In other words, classes may be merged (equal u_k^*) and/or removed (π_k equal to zero). To prevent local maxima this procedure should be repeated with several sets of starting values. Moreover, to guarantee that also the more difficult to find mass points located at $-\infty$ and $+\infty$ are encountered when needed in the NPML solution, it is a good idea to include latent classes located at $-\infty$ and $+\infty$ in each starting set (Hartzel *et al.*, 2001; Wood and Hinde, 1987). In the dichotomous response case we will deal with in the next sections, these correspond to success probabilities equal to 0 and 1, respectively.

2.3 Design of the simulation study

To keep the simulation study feasible, we will restrict ourselves to one particular type of GLMM, namely to the multilevel binary logistic regression model. The reason for this choice is that whereas binary outcome variables are very commonly used in sociological, behavioral, and biomedical studies, most attention is typically paid to models for continuous responses. Moreover, it is well documented that binary data are more sensitive to specification issues in multilevel analysis than continuous variables: in linear regression analysis, fully ignoring a random intercept does not bias parameter estimates, which is not the case in logistic regression analysis (Agresti *et al.*, 2000).

The population model we use is a two-level random-intercept logistic

regression model with one level-1 and one level-2 explanatory variable; that is,

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j. \tag{2.4}$$

We assume that x_{1ij} – the explanatory variable for level-1 unit i in level-2 unit j – takes on the values 0 and 1 with probability 0.5, and that x_{2j} – the explanatory variable for level-2 unit j – takes on the values 0 and 1 with probability 0.5 independently of x_{1ij}. For the fixed intercept β_0 and regression slopes β_1 and β_2, we used the same values across simulation replications. More specifically, we set their values to: $\beta_0 = -2$, $\beta_1 = \beta_2 = 2$. This yields large but not too extreme differences between the response probabilities for $u_j = 0$. More specifically, the corresponding response probabilities for the four possible combination of explanatory variables are

$$P(y = 1|x = 1, z = 1, u = 0) = e^2/(1 + e^2) = 0.88,$$

$$P(y = 1|x = 1, z = 0, u = 0) = e^0/(1 + e^0) = 0.5,$$

$$P(y = 1|x = 0, x = 1, u = 0) = e^0/(1 + e^0) = 0.5, \text{ and}$$

$$P(y = 1|x = 0, z = 0, u = 0) = e^{-2}/(1 + e^{-2}) = 0.12.$$

So far we discussed only the elements that were not varied in the simulations study. The factors that were varied are the specification of the random effects distribution and the level-1 and level-2 sample sizes. We wish to assess how the parametric and nonparametric models perform under different true random effects distributions and whether the performance depends on the level-1 and level-2 sample sizes.

Let us first look at the various specifications we used for the random effects distribution. We not only varied the form of the distribution, but also its variance. For the latter, it is important to note that in a logit model the level-1 errors are assumed to come from a logistic distribution with mean 0 and variance $\pi^2/3 \approx 3.29$. The ICC is therefore equal to:

$$ICC = \sigma_u^2/(\sigma_u^2 + 3.29). \tag{2.5}$$

Hox and Maas (2001) found that the value of the ICC may affect the accuracy of the estimates, which is why we included this factor in the simulation design. We set the ICC equal to 0.1 and 0.3, which corresponds to small and moderate values. The random effects variance σ_u^2 is easily derived from the above formula: $\sigma_u^2 = 3.29 \cdot ICC \,/\, (1 - ICC)$.

Data sets were generated using six distributional forms for the random effects, three continuous distributions – exponential, normal, and uniform – and three two-class discrete mixing distributions with membership probabilities of 0.10, 0.25, and 0.50 for the first class. With these choices we have apart from the normal distribution, distributions that considerably deviate from normal in terms of skewness, kurtosis, and discontinuity.

The other two factors that were varied are the level-1 and level-2 sample sizes. More specifically, for the number of level-2 units we used $n = 30$, 100, and 1000 and for the number of level-1 units $n_j = 3$, 10, and 50. These sample sizes were chosen to be in agreement with the simulation studies of Kreft and de Leeuw (1998) and Maas and Hox

(2004), and to cover the full range of small, moderate, and large sample sizes encountered in biomedical, behavioral, and social science research. For example, in family surveys and in panel studies the combination of $n = 1000$ and $n_j = 3$ is rather common. Moreover, according to Kreft and de Leeuw (1998), $n = 30$ is the minimum number of level-2 units required for a meaningful multilevel analysis with random effects models. In organizational surveys, it is common to have about as many as 50 level-1 units within each level-2 unit, mostly combined with 30 to 100 level-2 units.

Combining the 4 design factors – ICC value, distributional form, level-2 sample size, and level-1 sample size – yields a total of $2 \times 6 \times 3 \times 3 = 108$ conditions. We generated 1000 data sets for each of these conditions. For each simulated data set, the unknown model parameters were estimated using the parametric approach assuming that random effects come from a normal distribution and using the NPML approach.

2.4 Results of the simulation study

The aim of the simulation study was to determine the bias and relative efficiency of the parametric and nonparametric random effects approaches under different true random effects distributions and sample sizes. Let θ be one of the parameters of interest, in our case the fixed effects β_0, β_1, and β_2, and the standard deviation of the random effects distribution σ_u, which in the nonparametric case is computed from the nodes'

Table 2.1: Efficiency for the conditions $n = 1000$, $ICC = 0.3$, and $n_j = 50$ or 10

n_j	True distribution	Model	$\lvert\hat{\beta}_{0s} - \beta_0\rvert$	$\lvert\hat{\beta}_{1s} - \beta_1\rvert$	$\lvert\hat{\beta}_{2s} - \beta_2\rvert$	$\lvert\hat{\sigma}_s - \sigma\rvert$
50	Exponential	Normal	0.05	0.02	0.05	0.11
		Nonparametric	0.04	0.02	0.03	0.06
	Normal	Normal	0.04	0.02	0.06	0.02
		Nonparametric	0.04	0.02	0.06	0.02
	Uniform	Normal	0.04	0.02	0.06	0.02
		Nonparametric	0.04	0.02	0.05	0.02
	Discrete with $p(u_{01}) = 0.5$	Normal	0.09	0.02	0.06	0.16
		Nonparametric	0.03	0.02	0.02	0.04
	Discrete with $p(u_{01}) = 0.25$	Normal	0.07	0.02	0.06	0.02
		Nonparametric	0.03	0.02	0.02	0.02
	Discrete with $p(u_{01}) = 0.1$	Normal	0.04	0.02	0.07	0.05
		Nonparametric	0.03	0.02	0.02	0.01
10	Exponential	Normal	0.08	0.04	0.06	0.11
		Nonparametric	0.06	0.04	0.05	0.10
	Normal	Normal	0.05	0.04	0.06	0.04
		Nonparametric	0.05	0.04	0.06	0.03
	Uniform	Normal	0.05	0.04	0.07	0.06
		Nonparametric	0.05	0.04	0.07	0.03
	Discrete with $p(u_{01}) = 0.5$	Normal	0.13	0.04	0.05	0.23
		Nonparametric	0.05	0.04	0.04	0.06
	Discrete with $p(u_{01}) = 0.25$	Normal	0.07	0.04	0.06	0.04
		Nonparametric	0.05	0.04	0.05	0.03
	Discrete with $p(u_{01}) = 0.1$	Normal	0.06	0.04	0.08	0.11
		Nonparametric	0.05	0.04	0.06	0.02

locations and weights. The ML estimate of θ obtained in replication s, $s = 1, \ldots, 1000$, is denoted by $\widehat{\theta}_s$. Rather than using the standard definitions of bias and relative efficiency – $E(\widehat{\theta}_s - \theta)$ and $E\left[(\widehat{\theta}_s - \theta)^2\right]$ – we used a more robust definition to prevent that the results are affected by a very small number of replications with boundary estimates. More specifically, when using the NPML estimator, especially in the conditions with large number of level-2 units and small number of level-1 units, there is a (small) positive probability that one of the latent classes is located at

Table 2.2: Efficiency for the conditions $n_j = 3$, $ICC = 0.3$, and $n = 1000$, 100, or 30

n	True distribution	Model	$\lvert\hat{\beta}_{0s} - \beta_0\rvert$	$\lvert\hat{\beta}_{1s} - \beta_1\rvert$	$\lvert\hat{\beta}_{2s} - \beta_2\rvert$	$\lvert\hat{\sigma}_s - \sigma\rvert$
1000	Exponential	Normal	0.10	0.08	0.09	0.12
		Nonparametric	0.11	0.09	0.09	0.20
	Normal	Normal	0.07	0.08	0.10	0.07
		Nonparametric	0.08	0.08	0.10	0.11
	Uniform	Normal	0.08	0.08	0.10	0.08
		Nonparametric	0.09	0.08	0.10	0.08
	Discrete with $p(u_{01}) = 0.5$	Normal	0.11	0.07	0.08	0.17
		Nonparametric	0.11	0.08	0.07	0.21
	Discrete with $p(u_{01}) = 0.25$	Normal	0.08	0.08	0.09	0.07
		Nonparametric	0.09	0.08	0.09	0.07
	Discrete with $p(u_{01}) = 0.1$	Normal	0.08	0.08	0.10	0.06
		Nonparametric	0.09	0.08	0.10	0.10
100	Exponential	Normal	0.24	0.24	0.28	0.24
		Nonparametric	0.30	0.26	0.29	0.33
	Normal	Normal	0.26	0.25	0.29	0.21
		Nonparametric	0.28	0.27	0.28	0.26
	Uniform	Normal	0.24	0.25	0.29	0.21
		Nonparametric	0.29	0.27	0.31	0.24
	Discrete with $p(u_{01}) = 0.5$	Normal	0.25	0.23	0.27	0.33
		Nonparametric	0.29	0.24	0.26	0.54
	Discrete with $p(u_{01}) = 0.25$	Normal	0.25	0.25	0.30	0.22
		Nonparametric	0.29	0.25	0.30	0.25
	Discrete with $p(u_{01}) = 0.1$	Normal	0.25	0.25	0.31	0.22
		Nonparametric	0.28	0.26	0.32	0.22
30	Exponential	Normal	0.46	0.45	0.52	0.45
		Nonparametric	0.58	0.50	0.61	0.98
	Normal	Normal	0.46	0.47	0.54	0.42
		Nonparametric	0.59	0.54	0.64	0.68
	Uniform	Normal	0.46	0.51	0.54	0.41
		Nonparametric	0.63	0.58	0.65	0.80
	Discrete with $p(u_{01}) = 0.5$	Normal	0.45	0.47	0.54	0.57
		Nonparametric	0.58	0.53	0.59	1.19
	Discrete with $p(u_{01}) = 0.25$	Normal	0.46	0.49	0.55	0.44
		Nonparametric	0.56	0.54	0.64	0.73
	Discrete with $p(u_{01}) = 0.1$	Normal	0.46	0.48	0.58	0.42
		Nonparametric	0.61	0.54	0.65	0.59

Table 2.3: Bias for the conditions $n = 1000$, $ICC = 0.3$, and $n_j = 50$ or 10

n_j	True distribution	Model	$\hat{\beta}_{0s} - \beta_0$	$\hat{\beta}_{1s} - \beta_1$	$\hat{\beta}_{2s} - \beta_2$	$\hat{\sigma}_s - \sigma$
50	Exponential	Normal	-0.04	0.00	0.01	-0.11*
		Nonparametric	-0.02	0.00	0.00	-0.05
	Normal	Normal	0.00	0.00	0.00	-0.01
		Nonparametric	0.00	0.00	0.00	0.00
	Uniform	Normal	-0.01	0.00	0.01	0.02
		Nonparametric	0.00	0.00	0.00	0.00
	Discrete with $p(u_{01}) = 0.5$	Normal	0.09	0.01	0.04	-0.16*
		Nonparametric	0.00	0.00	0.00	0.00
	Discrete with $p\,(u_{01}) = 0.25$	Normal	0.06	0.00	0.02	0.02
		Nonparametric	0.00	0.00	0.00	0.00
	Discrete with $p\,(u_{01}) = 0.1$	Normal	-0.02	0.00	0.05	0.05
		Nonparametric	0.00	0.00	0.00	0.00
10	Exponential	Normal	-0.07	0.01	0.00	-0.11*
		Nonparametric	-0.04	0.01	0.01	-0.09*
	Normal	Normal	0.00	0.00	0.01	0.00
		Nonparametric	0.00	0.00	0.01	-0.01
	Uniform	Normal	-0.01	0.00	0.02	0.06
		Nonparametric	0.00	0.01	0.01	0.01
	Discrete with $p\,(u_{01}) = 0.5$	Normal	0.13*	0.00	0.02	-0.23*
		Nonparametric	0.01	0.01	0.01	-0.01
	Discrete with $p\,(u_{01}) = 0.25$	Normal	0.06	0.00	0.00	0.02
		Nonparametric	-0.01	0.01	0.01	0.01
	Discrete with $p\,(u_{01}) = 0.1$	Normal	-0.03	0.01	0.05	0.11*
		Nonparametric	-0.02	0.01	0.02	0.01

⋆ Cases with medians absolute value over 5%.

$-\infty$ or $+\infty$. When such boundary estimates may occur $E(\widehat{\theta}_s - \theta)$ and $E\left[(\widehat{\theta}_s - \theta)^2\right]$ do not exist. This not only to applies to σ_u, but also to β_0, β_1, and β_2. To prevent this problem from occurring we define bias as the median of $(\widehat{\theta}_s - \theta)$ and relative efficiency as the median of $|\widehat{\theta}_s - \theta|$. For similar approaches, see Agresti *et al.* (2004); Galindo-Garre *et al.* (2004).

Table 2.1 reports results on relative efficiency for a level-2 sample

Table 2.4: Bias for the conditions $n_j = 3$, $ICC = 0.3$, and $n = 1000$, 100, or 30

n	True distribution	Model	$\hat{\beta}_{0s} - \beta_0$	$\hat{\beta}_{1s} - \beta_1$	$\hat{\beta}_{2s} - \beta_2$	$\hat{\sigma}_s - \sigma$
1000	Exponential	Normal	-0.08	0.01	0.01	-0.11*
		Nonparametric	-0.08	0.02	0.02	-0.18*
	Normal	Normal	-0.01	0.01	0.01	0.00
		Nonparametric	-0.02	0.01	0.01	-0.07
	Uniform	Normal	-0.01	0.01	0.02	0.06
		Nonparametric	-0.01	0.01	0.03	0.03
	Discrete with $p(u_{01}) = 0.5$	Normal	0.02	0.01	0.01	-0.05
		Nonparametric	0.11*	0.02	0.03	-0.21*
	Discrete with $p(u_{01}) = 0.25$	Normal	0.07	0.00	0.01	0.01
		Nonparametric	0.00	0.02	0.03	-0.01
	Discrete with $p(u_{01}) = 0.1$	Normal	-0.02	0.02	0.02	0.01
		Nonparametric	-0.03	0.03	0.03	0.10*
100	Exponential	Normal	-0.07	0.02	0.00	-0.14*
		Nonparametric	-0.11*	0.06	0.06	-0.17*
	Normal	Normal	-0.02	0.03	0.03	-0.04
		Nonparametric	-0.06	0.07	0.10*	-0.06
	Uniform	Normal	-0.02	0.01	0.02	0.03
		Nonparametric	-0.07	0.05	0.10*	0.02
	Discrete with $p(u_{01}) = 0.5$	Normal	0.06	0.00	0.02	-0.29*
		Nonparametric	0.13*	0.08	0.10*	-0.34*
	Discrete with $p(u_{01}) = 0.25$	Normal	0.09	0.01	0.01	-0.04
		Nonparametric	-0.02	0.06	0.09	0.01
	Discrete with $p(u_{01}) = 0.1$	Normal	-0.02	0.03	0.04	0.06
		Nonparametric	-0.09	0.09	0.12*	0.09*
30	Exponential	Normal	-0.11*	0.05	0.01	-0.19*
		Nonparametric	-0.21*	0.19*	0.19*	-0.20*
	Normal	Normal	-0.02	0.04	0.07	-0.10*
		Nonparametric	-0.19*	0.19*	0.31*	-0.11*
	Uniform	Normal	-0.02	0.03	0.04	0.01
		Nonparametric	-0.19*	0.20*	0.24*	0.07
	Discrete with $p(u_{01}) = 0.5$	Normal	0.00	0.07	0.04	-0.48*
		Nonparametric	0.11*	0.24*	0.26*	-0.61*
	Discrete with $p(u_{01}) = 0.25$	Normal	0.09	0.08	0.02	-0.16*
		Nonparametric	-0.09	0.22*	0.25*	-0.08*
	Discrete with $p(u_{01}) = 0.1$	Normal	-0.05	0.04	0.05	0.00
		Nonparametric	-0.24*	0.18*	0.28*	0.10*

* Cases with medians absolute value over 5%.

size of 1000, level-1 sample sizes of 50 and 10, and $ICC = 0.3$. It can be observed, that the assumption of a normally distributed random intercept can give a moderate loss of efficiency compared to the NPML estimator when the true distribution of random intercept is continuous but not normal. On the other hand, when the true random intercept is normal, a nonparametric approach does not yield any loss of efficiency. In all situations with a discrete true distribution, we find a considerable loss of efficiency when a misspecified parametric model is used. Though details are not provided here, very similar results were obtained for the same level-1 and ICC conditions – thus with 50 and 10 level-1 units and $ICC = 0.3$ – but with the smaller numbers of 100 and 30 level-2 units.

There is no need to present all the details on the results for $ICC = 0.1$, the condition corresponding to a small level-2 variance, since these can easily be summarized. Irrespective of the level-1 and level-2 sample sizes and the form of the true random effects distribution, the parametric and nonparametric estimates are equally efficient. This holds even if the distribution of random effects is misspecified.

The efficiency estimates obtained with the smallest level-1 unit sample size ($n_j = 3$) and the largest ICC ($ICC = 0.3$) are reported in Table 2.2. Under these conditions, the parametric approach clearly outperforms the nonparametric approach. The former is more efficient irrespective of whether the true underlying distribution is misspecified or not. Even for the discrete true distributions, the parametric approach is the preferred one in terms of efficiency. The differences become larger as the level-2 sample size decreases and are larger for σ_u than for the β parameters.

The second evaluation criterion of interest is the bias in the parameter estimates. As was indicated above, we quantified bias as the median of the difference between estimated and true parameter value across simulation replications. Table 2.3 provides the estimated biases of the parameter estimates for a level-2 unit sample size of 1000, level-1 unit sample sizes of 50 or 10, and $ICC = 0.3$, and Table 2.4 for the 3 conditions with level-1 sample size of 3 and $ICC = 0.3$. Reported biases are marked by a "*" in these two tables when they are larger than 5% of true parameter value.

The conclusions that can be derived from Tables 2.3 and 2.4 are similar to those from Tables 2.1 and 2.2. Table 2.3 shows that the bias of the NPML estimator is negligible for all true distributions. The parametric estimator yields biased estimates for σ_u and β_0 when the true distribution is discrete or when the true distribution is continuous but asymmetric (exponential distribution). Although not reported here, very similar results were obtained when the number of level-2 units is decreased to 100 and 30 (and level-1 and ICC settings kept constant).

As was also the case for efficiency, in the $ICC = 0.1$ conditions, the parametric and nonparametric approach perform equally well in terms of bias (results are not listed here). All biases are negligible, irrespectively of whether the random effect distribution is correctly specified or not.

The results reported in Table 2.4 show that with a small level-1 sample size and large ICC, the nonparametric approach performs worse than the parametric one even when in the latter the underlying random intercept distribution is misspecified.

Table 2.5: Parameter estimates and log-likelihood values for the logistic regression models estimated with the 1988 Bangladesh Fertility Survey data set

	No random intercept		Parametric		Nonparametric	
	Coef	SE	Coef	SE	Coef	SE
Intercept	-1.568	0.126	-1.690	0.148	-1.664	.
No children	0.000	.	0.000	.		
1 child	1.059	0.152	1.109	0.158	1.100	0.159
2 children	1.288	0.167	1.377	0.175	1.368	0.176
3 or more children	1.216	0.171	1.346	0.180	1.327	0.181
Age	-0.024	0.008	-0.027	0.008	-0.026	0.008
Urban	0.797	0.105	0.732	0.120	0.719	0.122
Intercept Std. Dev.			0.464	0.079	0.472	.
ICC			0.061		0.063	
Log-likelihood	-1228.365		-1206.674		-1204.523	

2.5 Real data example

The use of logistic regression analysis with a random intercept is illustrated with a data set from the 1988 Bangladesh Fertility Survey (Huq and Cleland, 1990). It contains information on 1934 women who live in 60 areas of Bangladesh. It is a two-level data set: women are the level-1 units which are nested within living areas, the level-2 units. The dependent variable is the use of contraceptives yes/no. Since this is a binary response variable, it is natural to use a logit link function. Level-1 predictors are the woman's number of living children measured in four categories (no children, 1 child, 2 children, 3 or more children), and the woman's age (centered around the mean). The single level-2 predictor is type of region of residence (urban or rural). Number of living children is used as a categorical predictor, which "no children" as the reference

category.

Using the Latent GOLD 4.0 software (Vermunt and Magidson, 2005), we estimated a standard logistic regression model without random effects, as well as parametric and nonparametric random intercept models. For the nonparametric model we used the "zero-inflated" option to make sure that mass points at $-\infty$ and $+\infty$ are encountered. Table 2.5 reports the parameter estimates and the value of the log-likelihood function for the three estimated models. Comparison of the log-likelihood values indicates that the random intercept is needed. The nonparametric specification yields a slightly larger log-likelihood value than the parametric specification. The NPML solution contains 5 mass points which are located at -1.138, -2.342, -1.608, -1.867, and $-\infty$ with weight equal to 0.350, 0.255, 0.254, 0.1299, and 0.011, respectively.

The parameter estimates obtained with the parametric and nonparametric approach are very similar in this application. This confirms the results of our simulation study in which we found that the two approaches yield almost indistinguishable results for small ICC values (note that the ICC is about 0.06 in this application). It should be noted that we excluded the mass point located at $-\infty$ and with a very small weight in the computation of the mean and the standard deviation of the intercept for the NPML solution. Inclusion of this mass point yields a mean equal to $-\infty$ and a standard deviation equal to ∞.

2.6 Conclusions and discussion

The two questions that we wished to answer using the simulation study were 1) whether the NPML estimator performs better in terms of bias and efficiency compared the parametric model when the latter is misspecified, and 2) whether the NPML estimator performs equally well in terms of bias and efficiency compared the parametric model when the latter is correctly specified. This was studied for small and large level-1 and level-2 sample sizes, for small and moderate ICC values, and for different types of random effects distributions. We are now able to answer these two questions for the two-level random intercept logistic regression model.

The simulation study showed that the results depend strongly on the level-1 sample size and on the ICC values. More specifically for the larger ICC value and moderate or large level-1 sample size, we found exactly what we expected: the NPML method performs better than the parametric method when assumptions of the latter are violated and equally well when they are not violated. In such cases we should thus always use a NPML estimator since we do not know whether the assumptions hold. For small ICC values, both approaches perform equally well, so either of the two can be used in such situations. Again, it does not harm using the NPML method when the assumptions of the parametric approach hold.

In one set of conditions the NPML method turned out to be problematic; that is, when the number of level-1 units is very small ($n_j = 3$)

and the ICC is not very small ($ICC = 0.3$). In these conditions the parametric approach outperformed the NPML estimator even when the true underlying distribution of random intercept was far from normal. In other words, when the number of level-1 units is very small, it is better to use a parametric random effects model.

The results of our study are in agreement with the studies by Hartzel *et al.* (2001) and Agresti *et al.* (2004), which as mentioned in the introduction, yielded seemingly contradictory results. Similar to Hartzel *et al.* (2001), we found that with small level-1 sample sizes it may be better to use a parametric random effects model, even if this misspecifies the true random effects distribution. Moreover, similar to Agresti *et al.* (2004), we found that with moderate and large level-1 sample sizes and larger ICC values, using a nonparametric approach is preferred when the underlying assumptions of the parametric model do not hold and does not harm when they hold. In other words, the level-1 sample size and the ICC value are the critical factors.

One limitation of our study is that, as in the studies by Hartzel *et al.* (2001) and Agresti *et al.* (2004), we investigated the performance of the methods only in terms of bias and efficiency of the estimated fixed and random effects parameters. We are aware of the fact that sometimes prediction of the random effects may even be more important than estimation of the fixed and random effects parameters. Another simulation study would be needed to determine how well the various methods perform in terms of prediction.

Another limitation of our study is that it concerns logistic regression

models with only a random intercept. It is not clear whether our findings can be generalized to models containing also random slopes; that is, from the univariate to the multivariate random effects case. Random slopes introduce several additional complications, both in the parametric and nonparametric approach. In future research, we will investigate whether the conclusions drawn here also apply to models with random slopes.

In our study we investigated two different specifications for the random effects distribution: the parametric approach with an underlying normal distribution and the nonparametric approach using an unspecified discrete mixing distribution. As a third alternative one may use a combination of these two: a finite mixture of normal distributions (Magder and Zeger, 1996; Verbeke and Molenberghs, 2000). Whereas such an approach may have particular advantages, such as that contrary to the nonparametric approach it yields non-discrete random effects, Agresti *et al.* (2004) obtained somewhat disappointing results with this approach in the context of a log linear model for an odds ratio. Nevertheless, we believe that this hybrid approach may be promising in other situations.

Another topic that we did not address in this article is the possibility to use a semi-parametric approach in which the number of mass points is not increased till the maximum of the log-likelihood is found, but in which instead a penalized log-likelihood is maximized (or minimized). A possibility may, for example, be to select the number of mass points minimizing the Bayesian information criterion (BIC).

Chapter 3

Logistic regression analysis with multidimensional random effects: a comparison of three approaches

3.1 Introduction

During the last decades, multilevel regression analysis has become part of the standard statistical toolbox of researchers in the social and behavioral sciences as well as in the biomedical field. This statistical method is used for the analysis of data sets in which lower-level units are nested within higher-level units (Hox, 2002; Skrondal and Rabe-Hesketh, 2004; Snijders and Bosker, 1999). Examples include data sets with a nesting of persons within families, survey respondents within geographical units, patients within therapists, pupils within schools, employees within firms, and repeated measurements within subjects. The lower level of the hierarchical structure is often referred to as level-1 and the higher level as level-2.

Typical for multilevel data is that level-1 observations belonging to the same level-2 unit are more alike than level-1 units from different level-2 units, for example, because they share common environments, experiences, and interactions. The implications of this is that the responses of level-1 units within the same level-2 units are correlated and can thus not be treated as independent observations in the statistical analysis. Whereas in some applications this is perceived as a problem that should be dealt with when modeling multilevel data, in other applications the multilevel data structure is seen as containing valuable information on how groups (higher-level units) differ from each other, for example, in terms of the effects of explanatory variables on the outcome variable of interest (Bryk and Raudenbush, 1992; Hox, 1994; Snijders and Bosker, 1999).

The most popular approach for the analysis of such data sets is by means of multilevel models, which are also referred to as hierarchical, mixed, random-effects, or random-coefficients models (Bryk and Raudenbush, 1992; Hox, 1994; Longford, 1995; Snijders and Bosker, 1999). Whereas the terms "multilevel" and "hierarchical" refer to the data structure, the terms "mixed", "random-effects" and "random-coefficients" indicate what these models are from a more technical point of view. More specifically, these models capture differences between level-2 units – and thus also correlations between level-1 observations within level-2 units – by allowing one or more of the model parameters to vary randomly across level-2 units. Whereas the earliest developments and applications of multilevel regression models concerned linear models for continuous

responses, these are nowadays also applied with discrete response variables. The most popular model for binary responses is the random-effects logistic regression model (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993).

A key issue in the specification of a multilevel regression model is that not only assumptions have to be made about the distribution of the residuals, but also about the distribution of the random effects, also referred to as the mixing distribution. The most common approach is to assume that it has a convenient parametric form, in most cases a normal distribution. However, as stressed by Aitkin (1999), parametric distributional assumptions about the random effects will usually not hold in practice, which may have serious implications for the parameter estimates. For example, various studies found that misspecification of the distribution of random effects results in a loss of efficiency of the fixed coefficient estimates (Agresti *et al.*, 2004; Heagerty and Kurland, 2001; Maas and Hox, 2004; Neuhaus *et al.*, 1992). Lukočienė and Vermunt (2008) not only confirmed this result for the random-intercept logistic regression model, but also showed that the estimate for the random-intercept variance may be severely biased when its distribution is misspecified.

Rather than using a parametric random-effects approach, it is also possible to use either a nonparametric or a semi-parametric approach. These two alternatives have in common that they are both latent class models; that is, a discrete mixing distribution with K nodes (latent classes) is used to approximate the underlying distribution with an unknown shape. The locations and weights corresponding to the nodes

are quantities to be estimated. Although the nonparametric and semi-parametric approach are similar, they are fundamentally different in how they determine the number of latent classes. In the former, the number of latent classes is increased till the likelihood function is maximized, which yields what is called the nonparametric maximum likelihood (NPML) estimator of the random-effects distribution (Heckman and Singer, 1984; Laird, 1978; Lindsay, 1995). As indicated by Leroux (1992*b*) and Leroux and Puterman (1992), the NPML estimate may yield unnecessarily large numbers of latent classes and well fitting models with fewer latent classes may be preferred. Rather than increasing the number of latent classes till a saturation point is reached, it is also possible to decide about the number of classes using information criteria such as AIC and BIC. Note that this is what is usually done in mixture regression analysis (Vermunt and Dijk, 2001; Wedel and DeSarbo, 1994), as well as in other types of latent class analyses. To distinguish this approach from NPML, we call it a semi-parametric approach.

This paper provides a comparison of the three random-effect approaches within the context of multilevel logistic regression analysis. It extends the work by (Lukočienė and Vermunt, 2008) on the comparison of parametric and NPML approaches for random-effects logistic regression analysis to the situation in which not only the intercept but also slopes are random coefficients, as is usual in social and behavioral science application of multilevel regression analysis (Kreft and de Leeuw, 1998; Singer, 1998; Snijders and Bosker, 1999). As far as we know, there are no studies investigating the performance of the NMPL approach when

applied with multidimensional random effects. Moreover, we include the semi-parametric approach in the comparison. This is the commonly used latent-class based regression modeling approach for situations in which not only the intercept but also the slopes vary randomly across level-2 units. We focus on binary logistic regression models because these are more sensitive to specification issues in multilevel analysis than models for continuous response variables or counts (Agresti *et al.*, 2000, 2004).

Using a simulation study we wish to find out which of the three approaches – parametric, semi-parametric or nonparametric – should be used under different types of true random-effects distributions and specific features of the sample. More specifically, we are interested in whether it makes sense to use a nonparametric or semi-parametric model as an alternative when the underlying assumptions of the parametric model do not hold? Moreover, we wish to know whether it harms to use a nonparametric or semi-parametric model – say for practical reasons – when the assumptions of the parametric model hold?

The next section describes the multilevel logistic regression model of interest. Section 3.3 discusses the set up of the simulation study. Results of the simulation study are presented in Section 3.4. The last section summarizes the main conclusions and provides some practical recommendations.

3.2 The two-level logistic regression model

This section describes the two-level logistic regression model using the single equation mixed model formulation (Skrondal and Rabe-Hesketh, 2004). An alternative would be to use the hierarchical model formulation, which contains separate regression equations for the various hierarchical levels (Bryk and Raudenbush, 1992; Hox, 2002; Snijders and Bosker, 1999).

Let y_{ij} denote the binary response ($y_{ij} = 0, 1$) of the level-1 unit i, $i = 1, \ldots, n_j$, belonging to the level-2 unit j, $j = 1, \ldots, n$. Explanatory variables are referred to by $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$, where the former concern the fixed and the later the random effects. The vector with fixed effects is denoted by $\boldsymbol{\beta}$ and the vector with the unobservable common random coefficients shared by all level-1 units belonging to the $j^{\underline{th}}$ level-2 unit by $\boldsymbol{u}_j$. Let $\pi_{ij} = E(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_j)$ be the conditional expectation of y_{ij}. The multilevel logistic regression model for y_{ij} takes on the following form:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \boldsymbol{\beta}' \boldsymbol{x}_{ij} + \boldsymbol{u}_j' \boldsymbol{z}_{ij}. \tag{3.1}$$

The typical assumption for the random coefficients $\boldsymbol{u}_j$ is that these are independently and identically distributed multivariate normal random variables with zero means and covariance matrix Σ_u. Consistent with this distributional assumption, parameters of the two-level logistic regression model may be estimated by maximum likelihood (ML), where construction of the likelihood function is simplified by the fact that the y_{ij} can be assumed to be independent within level-2 units conditionally

on the observed predictors and the unobserved random effects. ML estimation involves maximizing the following marginal likelihood function:

$$L(\boldsymbol{\beta}, \Sigma_u) = \prod_{j=1}^{n} \int_{\boldsymbol{u}_j} \left[\prod_{i=1}^{n_j} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \right] f(\boldsymbol{u}_j; \Sigma_u) d\boldsymbol{u}_j, \tag{3.2}$$

where $\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$ represents the Bernoulli distribution for the level-1 errors. Note that the fixed effects $\boldsymbol{\beta}$ and covariance matrix Σ_u are the unknown parameters to be estimated. The integral should be solve numerically, for example, using Gauss-Hermite quadrature, which is basically a discrete approximation of the multivariate normal integral. Algorithms for maximizing the resulting numerically integrated marginal likelihood are the EM algorithm (Agresti *et al.*, 2000; Bock and Aitkin, 1981; Dempster *et al.*, 1977) and gradient methods, such as the Fisher scoring (Longford, 1987) and Newton-Raphson algorithm (Pan and Thompson, 2003; Rabe-Hesketh *et al.*, 2004). In our study, we used numerical integration with 50 nodes per dimension. For maximization a combination of EM and Newton-Raphson was used, where the estimation process starts with EM iterations and switches to Newton-Raphson when the relative change in parameters is very small (Vermunt and Magidson, 2005).

As was indicated in the introduction, usually nothing or very little it is known about the underlying distribution of the random effects (Aitkin, 1999). To prevent possible misspecification, it may therefore be attractive to assume the random effects $\boldsymbol{u}_j$ come from an unspecified mixing distribution concentrated on a finite number of latent classes or mass points (Aitkin, 1999; Heckman and Singer, 1984; Laird, 1978; Vermunt,

1997). Let K denote the number of latent classes, k a particular latent class, and $\boldsymbol{u}_k^*$ the unknown values of the random effects $\boldsymbol{u}_j$ when level-2 unit j belongs to latent class k, and let $\pi_k = P(\boldsymbol{u}_j = \boldsymbol{u}_k^*)$ represent the probability that a randomly selected level-2 unit belongs to latent class k or in other words that the random effects correspond to the location of class k. Using such a K-class discrete characterization of the random effects distribution yields the following marginal likelihood function:

$$L(\boldsymbol{\beta}, \boldsymbol{u}^*, \boldsymbol{\pi}) = \prod_{j=1}^{n} \sum_{k=1}^{K} \left[\prod_{i=1}^{n_j} \pi_{ij|k}^{y_{ij}} (1 - \pi_{ij|k})^{1-y_{ij}} \right] \pi_k, \tag{3.3}$$

where $\pi_{ij|k}$ is the conditional density function of y_{ij} given that level-2 unit j belongs to latent class k. The two-level logistic regression model can now be written as a model for $\pi_{ij|k}$; that is,

$$\log \frac{\pi_{ij|k}}{1 - \pi_{ij|k}} = \boldsymbol{\beta}' \boldsymbol{x}_{ij} + \boldsymbol{u}_k^{*\prime} \boldsymbol{z}_{ij}. \tag{3.4}$$

The weights are restricted such that $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. In addition one identifying location constraint should be imposed on each of the $M + 1$ random coefficients, e.g. $\sum_{k=1}^{K} u_{mk}^* \pi_k = 0$, which implies that the $\boldsymbol{u}_k^* = (u_{0k}^*, \ldots, u_{mk}^*, \ldots, u_{Mk}^*)$ are centered. The unknown parameters to be estimated are the fixed effects $\boldsymbol{\beta}$, $K - 1$ free mass point locations per dimension $(u_{0k}^*, \ldots, u_{Mk}^*)$ and $K-1$ free mass point weights π_k. Note that although the variances and covariance of the random effects are not model parameters, they can easily be estimated as follows (Vermunt and Dijk, 2001):

$$\hat{\sigma}_m^2 = \sum_{k=1}^{K} (u_{mk}^*)^2 \pi_k \quad \text{and} \quad \hat{\sigma}_{mm'} = \sum_{k=1}^{K} u_{mk}^* u_{m'k}^* \pi_k.$$

Maximization of the marginal likelihood function in equation (3) for a specific K, as in the parametric case, can be achieved by means of the EM and/or Newton-Raphson algorithm. It is usually advised to use of multiple sets of starting values to reduce the likelihood of ending up in a local maximum.

To obtain the solution corresponding to the NPML estimate of the random effects distribution, we not only have to maximize (3) for specific values of K; but we simultaneously have to find the value of K – say K_{NPML} – that yields the largest marginal likelihood value. In other words, we have to find the saturation point at which increasing K no longer results in an increase of the likelihood function. A method to find K_{NPML} proposed by various authors involves introducing latent classes one by one using directional (Gateaux) derivatives (Böhning, 2000; Lindsay, 1983, 1995; Rabe-Hesketh *et al.*, 2003). A much simpler alternative approach is to estimate the model with a large number of latent classes, K_{MAX}. When $K_{MAX} > K_{NPML}$, the ML estimates for $\boldsymbol{u}_k^*$ will be equal for some latent classes and/or the estimate for π_k will be equal to zero for some latent classes (Böhning, 2000). In other words, classes may be merged (equal $\boldsymbol{u}_k^*$) and/or removed (π_k equal to zero). To prevent local maxima this procedure should be repeated with several sets of starting values. Moreover, to guarantee that also the more difficult to find mass points located at $-\infty$ and $+\infty$ are encountered when needed in the NPML solution, it is advisable to include latent classes located at

these values in each starting set (Hartzel *et al.*, 2001; Wood and Hinde, 1987).

As already mentioned, the NPML estimates may yield unnecessarily large K (Leroux, 1992*b*; Leroux and Puterman, 1992) and estimates with a smaller number of latent classes that describe the data sufficiently may be preferred. Moreover, the latent classes may have substantive interpretations which are useful for the study concerned. This yields an approach in which the value of K should be estimated, yielding what we called the semi-parametric random-effects modeling approach. In this approach the value of K is increased till the criterion used for model selection no longer improves. In our study, we will use the BIC (Schwarz, 1978) for deciding about the number of classes, as was for example done by Vermunt and Dijk (2001); Wedel and DeSarbo (1994).

3.3 Design of the simulation study

This section describes the design of the simulation study. First, we discuss the design factors that were kept constant, and subsequently the ones that were varied. The key factors that were kept contant are the overall structure of the population model, the values of the fixed-effect parameters, and the values of the intraclass correlations for the random effects.

The population model we used is a two-level random coefficients logistic regression model with one level-1 and one level-2 explanatory variable.

This model can be formulated as follows:

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_{0j} + u_{1j} z_{1ij}. \tag{3.5}$$

Here, both x_{1ij} and z_{1ij} represent the level-1 predictor (in fact, $x_{1ij} = z_{1ij}$), where x_{1ij} is used to define its fixed part and z_{1ij} its random part. The other fixed effects correspond to the intercept and the level-2 predictor x_{2j}. The two explanatory variables are assumed to be binary predictors taking on the values 0 and 1 with probability 0.5 independently of one another. For the fixed intercept β_0 and slopes β_1 and β_2, we used the same values across simulation replications. More specifically, we set their values to: $\beta_0 = -2$, $\beta_1 = \beta_2 = 2$. This yields large enough but not too extreme differences between the response probabilities for $u_{0j} = 0$ and $u_{1j} = 0$. More specifically, the corresponding response probabilities for the four possible combinations of explanatory variables are

$$P(y = 1|x_1 = 1, x_2 = 1, u_0 = u_1 = 0) = e^2/(1 + e^2) = 0.88,$$

$$P(y = 1|x_1 = 1, x_2 = 0, u_0 = u_1 = 0) = e^0/(1 + e^0) = 0.5,$$

$$P(y = 1|x_1 = 0, x_2 = 1, u_0 = u_1 = 0) = e^0/(1 + e^0) = 0.5, \text{ and}$$

$$P(y = 1|x_1 = 0, x_2 = 0, u_0 = u_1 = 0) = e^{-2}/(1 + e^{-2}) = 0.12.$$

A second element that was kept constant in the simulation study is the overall importance of the random part, which can be expressed by means of the intraclass correlation (ICC). Although Hox and Maas (2001) found that the value of the ICC may affect the impact of a misspecification of the random effects distribution, the study by Lukočienė

and Vermunt (2008) on the random-intercept logistic regression model found that parametric and nonparametric approaches are almost indistinguible when the ICC value is small (e.g. 0.1). We will therefore not investigate this situation again, but instead focus on the condition with a moderate ICC value of 0.3. For this ICC value, Lukočienė and Vermunt (2008) found important differences in the performance of the parametric and nonparametric approaches.

The ICC values can be set by using the fact that level-1 errors coming from a logistic distribution have a variance equal to $\pi^2/3$. Since $ICC = \sigma^2/(\sigma^2 + \pi^2/3)$, the variance of the random intercept σ_0^2 can be obtained by $\sigma_0^2 = ICC\,/\,(1 - ICC)\,\pi^2/3$, which for $ICC = 0.3$ yields $\sigma_0^2 = 1.41$. Similar to Busing (1993) and Maas and Hox (2004), we used the same variance for the random slope as for the random intercept ($\sigma_1^2 = 1.41$ as well).

So far, we discussed only the elements that were not varied in the simulations study. The three design factors that were varied are the random effects distribution, the level-1 sample size, and the level-2 sample size. We wish to assess how the parametric, nonparametric, and semi-parametric models perform under different true random-effect distributions and whether the performance depends on the level-1 and level-2 sample sizes. The study by Lukočienė and Vermunt (2008) on the random-intercept logistic regression model showed that these are the main factors affecting the performance of the parametric and nonparametric approaches.

Data sets were generated using four distributional forms for the ran-

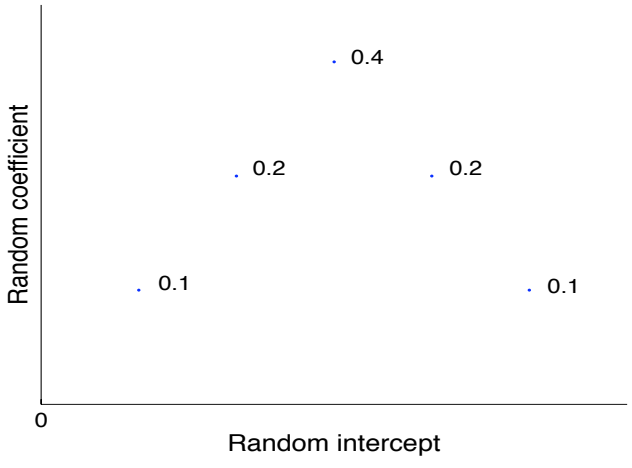

Figure 3.1: Discrete mixing distribution with five classes

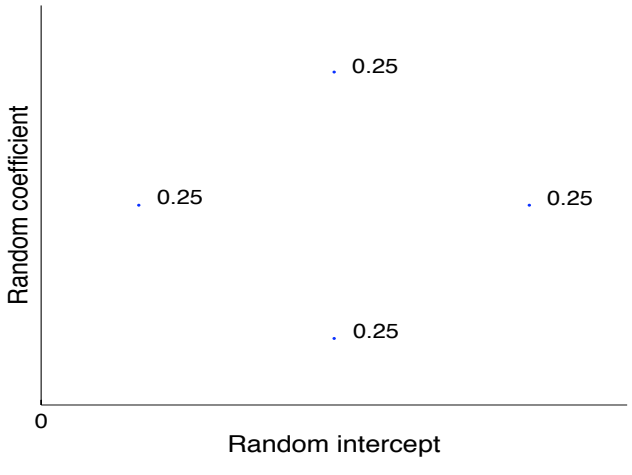

Figure 3.2: Discrete mixing distribution with four classes

dom effects, two continuous distributions (exponential and normal) and two discrete mixing distributions – one five-class distribution with class membership probabilities of 0.1, 0.2, 0.4, 0.2, and 0.1, respectively, and another four-class distribution with equal membership probabilities of 0.25. As demonstrated in Figure 3.1 and Figure 3.2, the locations of the classes of these two discrete mixing distribution were chosen in such a way that the random intercept and random slope would be uncorrelated, but strongly associated. With these four choices we have apart from the

normal distribution, distributions that considerably deviate from normal in terms of skewness, kurtosis, discontinuity, and association between dimensions.

The other two factors that were varied are the level-1 and level-2 sample sizes. More specifically, for the number of level-2 units we used $n = 30$, 100, and 1000 and for the number of level-1 units $n_j = 10$, and 50. These sample sizes reflect the typical sample sizes in multilevel analysis (see also Kreft and de Leeuw (1998); Lukočienė and Vermunt (2008); Maas and Hox (2004)).

Combining the 3 design factors – distributional form, level-2 sample size, and level-1 sample size – yielded a total of $4 \times 2 \times 3 = 24$ conditions. We generated 1000 simulated data sets for each of these conditions. For each simulated data set, the unknown model parameters were estimated using the parametric approach assuming that random effects come from a normal distribution, the NPML approach, and the semi-parametric approach using BIC as the model selection criterion.

3.4 Results of the simulation study

The aim of the simulation study was to determine the bias and relative efficiency of the parametric, nonparametric, and semi-parametric random effects approaches under the different true random effects distributions and sample sizes. Let θ be one of the parameters of interest, which in our case are the fixed effects β_0, β_1, and β_2, and the standard deviations of the random effects distribution σ_0 and σ_1 which in the nonparametric

and semi-parametric cases are computed from the nodes' locations and weights. The ML estimate of θ obtained in replication s, $s = 1, \ldots, 1000$, is denoted by $\widehat{\theta}_s$. Rather than using the more standard definitions of bias and relative efficiency – $E(\widehat{\theta}_s - \theta)$ and $E\left[(\widehat{\theta}_s - \theta)^2\right]$ – we used a more robust definition to prevent that the results are affected by a small number of replications with boundary estimates. More specifically, when using the NPML estimator, especially in the conditions with large number of level-2 units and small number of level-1 units, there is a positive probability that one of the latent classes is located at infinity. In our case, latent classes can have 4 such possible locations: $(-\infty, -\infty)$, $(-\infty, \infty)$, (∞, ∞), and/or $(\infty, -\infty)$. When such boundary estimates may occur $E(\widehat{\theta}_s - \theta)$ and $E\left[(\widehat{\theta}_s - \theta)^2\right]$ do not exist. This not only applies to σ_0 and σ_1, but also to β_0, β_1, and β_2. To prevent this problem from occurring we define bias as the median of $(\widehat{\theta}_s - \theta)$ and relative efficiency as the median of $|\widehat{\theta}_s - \theta|$. For similar approaches, see Agresti *et al.* (2004) and Galindo-Garre *et al.* (2004).

Below we first discuss the results for the fixed effects and then for the random effects.

3.4.1 Fixed effects

The first evaluation criterion of interest is the bias in the parameter estimates. Table 3.1 provides the estimated biases of the fixed effects for the level-2 sample sizes of 1000 and 30. The first three columns of Table 3.1 indicate the values for the design factors: level-2 sample size, level-1 sample size, and the random-effects distribution used to

Table 3.1: Bias of the fixed effects for the conditions $n = 1000$ and 30

n	n_j	True distribution	Assumed	$\hat{\beta}_{0s} - \beta_0$	$\hat{\beta}_{1s} - \beta_1$	$\hat{\beta}_{2s} - \beta_2$
1000	10	Exponential	Normal	-0.12*	-0.03	0.04
			Nonparametric	-0.09	0.90*	0.04
			Semi-parametric	0.04	-0.09	-0.08
		Normal	Normal	0.00	0.01	0.03
			Nonparametric	-0.17*	0.28*	0.05
			Semi-parametric	0.06	-0.07	-0.06
		Discrete (4 classes)	Normal	-0.02	0.00	0.04
			Nonparametric	-0.11*	0.24*	0.05
			Semi-parametric	0.00	-0.01	-0.01
		Discrete (5 classes)	Normal	0.07	-0.21*	0.02
			Nonparametric	-0.02	0.13*	0.01
			Semi-parametric	0.00	-0.01	0.00
1000	50	Exponential	Normal	-0.04	-0.05	-0.03
			Nonparametric	-0.01	0.40*	0.01
			Semi-parametric	0.04	-0.04	-0.05
		Normal	Normal	0.00	0.00	0.02
			Nonparametric	-0.01	0.15*	0.03
			Semi-parametric	0.08	-0.03	-0.11*
		Discrete (4 classes)	Normal	-0.06	0.04	-0.24*
			Nonparametric	-0.01	0.13*	0.02
			Semi-parametric	0.00	0.00	0.01
		Discrete (5 classes)	Normal	0.05	-0.21*	0.05
			Nonparametric	-0.02	0.16*	0.01
			Semi-parametric	-0.01	0.00	0.00
30	10	Exponential	Normal	-0.10*	-0.02	0.01
			Nonparametric	-0.58*	2.35*	0.29*
			Semi-parametric	0.06	0.03	-0.06
		Normal	Normal	-0.02	0.07	0.02
			Nonparametric	-0.94*	2.46*	0.32*
			Semi-parametric	0.096	0.01	-0.18*
		Discrete (4 classes)	Normal	0.02	0.05	-0.03
			Nonparametric	-0.88*	2.19*	0.28*
			Semi-parametric	0.08	-0.02	-0.04
		Discrete (5 classes)	Normal	0.105*	-0.16*	-0.08
			Nonparametric	-0.16*	2.09*	0.02
			Semi-parametric	0.01	0.11*	0.00
30	50	Exponential	Normal	-0.06	-0.04	-0.08
			Nonparametric	-0.24*	1.59*	-0.14*
			Semi-parametric	0.05	-0.03	-0.01
		Normal	Normal	0.00	0.04	-0.09
			Nonparametric	-0.26*	0.92*	-0.34*
			Semi-parametric	0.07	0.01	-0.16*
		Discrete (4 classes)	Normal	-0.07	0.07	-0.15*
			Nonparametric	-0.15*	0.22*	0.04
			Semi-parametric	0.00	0.01	0.01
		Discrete (5 classes)	Normal	0.10*	-0.16*	-0.08
			Nonparametric	-0.17*	2.20*	0.01
			Semi-parametric	0.01	0.14*	0.00

* Cases with medians absolute value over 5%.

generate the data sets. The fourth column indicates which of the three approaches – parametric, nonparametric or semi-parametric – was used

Table 3.2: Efficiency of the fixed effects for the conditions $n = 1000$ and30

n	n_j	True distribution	Assumed	$\lvert\hat{\beta}_{0s} - \beta_0\rvert$	$\lvert\hat{\beta}_{1s} - \beta_1\rvert$	$\lvert\hat{\beta}_{2s} - \beta_2\rvert$
1000	10	Exponential	Normal	0.12	0.05	0.067
			Nonparametric	0.13	0.93	0.08
			Semi-parametric	0.05	0.11	0.070
		Normal	Normal	0.06	0.05	0.06
			Nonparametric	0.18	0.87	0.070
			Semi-parametric	0.07	0.07	0.071
		Discrete (4 classes)	Normal	0.06	0.07	0.09
			Nonparametric	0.11	0.30	0.08
			Semi-parametric	0.04	0.05	0.07
		Discrete (5 classes)	Normal	0.08	0.21	0.10
			Nonparametric	0.04	0.15	0.023
			Semi-parametric	0.03	0.04	0.021
1000	50	Exponential	Normal	0.06	0.05	0.07
			Nonparametric	0.04	0.40	0.04
			Semi-parametric	0.05	0.06	0.06
		Normal	Normal	0.04	0.04	0.07
			Nonparametric	0.05	0.23	0.09
			Semi-parametric	0.09	0.05	0.13
		Discrete (4 classes)	Normal	0.06	0.05	0.24
			Nonparametric	0.03	0.04	0.03
			Semi-parametric	0.02	0.02	0.02
		Discrete (5 classes)	Normal	0.07	0.21	0.10
			Nonparametric	0.04	0.16	0.021
			Semi-parametric	0.03	0.03	0.020
30	10	Exponential	Normal	0.29	0.27	0.388
			Nonparametric	0.80	3.07	0.50
			Semi-parametric	0.32	0.39	0.391
		Normal	Normal	0.29	0.29	0.38
			Nonparametric	1.08	3.94	0.59
			Semi-parametric	0.36	0.34	0.47
		Discrete (4 classes)	Normal	0.32	0.322	0.44
			Nonparametric	0.96	3.51	0.51
			Semi-parametric	0.31	0.320	0.36
		Discrete (5 classes)	Normal	0.27	0.23	0.35
			Nonparametric	0.28	2.13	0.16
			Semi-parametric	0.21	0.44	0.14
30	50	Exponential	Normal	0.27	0.20	0.36
			Nonparametric	0.37	1.85	0.38
			Semi-parametric	0.35	0.38	0.35
		Normal	Normal	0.27	0.21	0.41
			Nonparametric	0.45	1.64	0.58
			Semi-parametric	0.30	0.23	0.48
		Discrete (4 classes)	Normal	0.29	0.23	0.37
			Nonparametric	0.27	0.44	0.15
			Semi-parametric	0.18	0.18	0.14
		Discrete (5 classes)	Normal	0.28	0.23	0.34
			Nonparametric	0.29	2.22	0.16
			Semi-parametric	0.21	0.46	0.14

for the estimation of the parameters. The last three columns present the biases in the estimates of the intercept and the two slopes. Reported biases are marked by a "*" when they are larger than 5% of the true

parameter value, and smaller values are considered negligible.

Table 3.1 shows that the bias of the fixed effects in the models estimated under the semi-parametric approach is negligible for all cases with true discrete and exponential underlying distributions, except for the fixed effect β_1 when $n = 30$ and true underlying discrete distribution has 5 latent classes. When the semi-parametric approach is applied with the true bivariate normal distribution, we see biases only for β_2. The NPML approach yields biased estimates for almost every parameter, where it makes no difference whether the true distribution is discrete or continuous. The parametric approach performs very well when the true underlying distribution is bivariate normal, in which case the bias in the fixed effects is always negligible. However, for other true underlying distributions, the parametric approach gives biased estimates for at least one of the fixed effects. The results for the medium level-2 unit sample size $n = 100$ (which are not shown) are very similar to the results obtained with $n = 1000$.

As was mentioned above, the second evaluation criterion of interest is the efficiency of the parameter estimates. Table 3.2 reports results on relative efficiency of the fixed effects obtained with the largest and smallest level-2 unit sample sizes $n = 1000$ and $n = 30$ (results for $n = 100$ are again similar to the ones for $n = 1000$). Table 3.2 can be read similarly to Table 3.1.

The semi-parametric approach clearly outperforms the parametric and nonparametric approaches in cases when the true underlying distribution of random effects is discrete. However, when the true un-

derlying random effects distribution is bivariate normal the parametric approach is most efficient. For the true underlying exponential distribution, the parametric and semi-parametric approaches perform equally well in terms of efficiency. We find a considerable lower efficiency under the nonparametric approach for almost every condition.

If we have a closer look at results presented in Table 3.1 and Table 3.2 from the perspective of the effects of the level-1 and level-2 sample sizes, it can be observed is that the nonparametric approach perform very bad with the smaller level-2 sample size, and this is enforced when also the level-1 sample size is small. The quality of the other two approaches is less strongly affected by the sample sizes. However, when misspecified, the normal model performs worse when the level-1 sample size increases.

3.4.2 Random effects

Table 3.3 shows the results on bias and relative efficiency for the random effects obtained with sample sizes $n = 1000$ and $n = 30$. As in Table 3.1, biases larger than 5% of the true parameter value are marked by a "*". The semi-parametric approach yields negligible biases for both random effects when the true underlying distribution is discrete and the sample size is 1000. The parametric approach yields moderate biases for almost every condition. However, the obtained biases of the parametric estimates with true continuous underlying distributions in the smallest samples ($n = 30$ and $n_j = 10$) are smaller than for the semi-parametric and nonparametric estimates. In most other cases, the semi-parametric approach performs best showing the smallest bias for all true distribu-

Table 3.3: Bias and efficiency of the random effects for the conditions $n = 1000$ and 30

n	n_j	True distribution	Assumed	$\hat{\sigma}_{0s} - \sigma_0$	$\hat{\sigma}_{1s} - \sigma_1$	$\lvert\hat{\sigma}_{0s} - \sigma_0\rvert$	$\lvert\hat{\sigma}_{1s} - \sigma_1\rvert$
1000	10	Exponential	Normal	0.35*	-0.17*	0.35	0.17
			Nonparametric	0.57*	6.15*	0.57	6.15
			Semi-parametric	-0.12*	-0.26*	0.13	0.32
		Normal	Normal	0.26*	0.28*	0.26	0.28
			Nonparametric	1.20*	6.59*	1.20	6.59
			Semi-parametric	-0.09*	-0.17*	0.10	0.18
		Discrete (4 classes)	Normal	0.29*	0.74*	0.29	0.74
			Nonparametric	0.16*	2.26*	0.16	2.26
			Semi-parametric	0.02	0.00	0.07	0.07
		Discrete (5 classes)	Normal	-0.02	-0.39*	0.07	0.39
			Nonparametric	0.03	2.25*	0.03	2.25
			Semi-parametric	-0.01	0.00	0.03	0.04
1000	50	Exponential	Normal	0.13*	-0.26*	0.13	0.26
			Nonparametric	-0.02	3.30*	0.07	3.30
			Semi-parametric	-0.05	-0.13*	0.06	0.16
		Normal	Normal	0.11*	0.18*	0.11	0.18
			Nonparametric	0.02	1.90*	0.05	1.90
			Semi-parametric	-0.05	-0.08*	0.04	0.08
		Discrete (4 classes)	Normal	0.24*	0.89*	0.24	0.89
			Nonparametric	0.02	0.09*	0.03	0.09
			Semi-parametric	0.00	0.01	0.02	0.02
		Discrete (5 classes)	Normal	-0.04	-0.39*	0.08	0.39
			Nonparametric	0.01	2.81*	0.04	2.81
			Semi-parametric	-0.01	0.00	0.03	0.04
30	10	Exponential	Normal	0.15*	-0.23*	0.61	0.81
			Nonparametric	2.45*	9.10*	2.45	9.10
			Semi-parametric	-0.18*	-0.47*	0.35	0.88
		Normal	Normal	0.05	0.29*	0.54	0.83
			Nonparametric	3.40*	10.69*	3.40	10.69
			Semi-parametric	-0.22*	-0.42*	0.39	0.75
		Discrete (4 classes)	Normal	0.14*	0.90*	0.57	1.06
			Nonparametric	2.78*	9.50*	2.78	9.50
			Semi-parametric	-0.27*	-0.33*	0.37	0.48
		Discrete (5 classes)	Normal	0.23*	-0.40*	0.42	0.52
			Nonparametric	0.18*	7.75*	0.24	7.75
			Semi-parametric	-0.05	0.12*	0.16	0.51
30	50	Exponential	Normal	0.25*	-0.24*	0.49	0.50
			Nonparametric	0.39*	6.87*	0.51	6.69
			Semi-parametric	-0.11*	-0.13*	0.26	0.47
		Normal	Normal	0.23*	0.39*	0.39	0.55
			Nonparametric	0.95*	6.35*	0.95	6.35
			Semi-parametric	-0.08*	-0.10*	0.18	0.28
		Discrete (4 classes)	Normal	0.37*	0.90*	0.45	0.91
			Nonparametric	0.14*	0.72*	0.21	0.72
			Semi-parametric	-0.05	0.00	0.15	0.14
		Discrete (5 classes)	Normal	0.22*	-0.41*	0.41	0.65
			Nonparametric	0.21*	7.93*	0.26	7.93
			Semi-parametric	-0.04	0.14*	0.16	0.51

* Cases with medians absolute value over 5%.

tions.

The last three columns of Table 3.3 report the information on the

efficiency of the random effects estimates (of the standard deviations of the random effects). For the random intercept, the semi-parametric approach outperforms the parametric and nonparametric approaches in all investigated conditions. The same applies to the random slope, except for one situation; that is, when $n_j = 10$ and the true underlying distribution is exponential, the parametric approach is the most efficient method.

Results on bias and relative efficiency of random effects for the medium level-2 sample size ($n = 100$) are again not presented because they are rather similar to the results obtained with $n = 1000$.

3.4.3 Remarks on semi-parametric and nonparametric approaches

The results reported in Tables 3.1, 3.2, and 3.3 show that the nonparametric approach performs worse than the semi-parametric approach in almost all investigated conditions. As explained earlier, the difference between these two approaches is that they use different methods for determining the number of latent classes. To see how the use of BIC worked out in our simulation study, let us take a look at the number of latent classes selected according to this criterion when the true distribution is discrete. More specifically, Table 3.4 presents the percentage of simulation replications (out of 1000) in which a particular number of latent classes was selected using the semi-parametric approach. As can be seen, the number of latent classes is often underestimated with the smaller level-2 sizes, and this tendency is stronger when also the level-1 sample size is small. It can also be observed that the semi-parametric

Table 3.4: Percentage of replications selecting a particular number of latent classes based on BIC in semi-parametric approach

n	n_j	True distribution	2 classes	3 classes	4 classes	5 classes
1000	50	Discrete with 4 classes			100	
		Discrete with 5 classes				100
	10	Discrete with 4 classes			100	
		Discrete with 5 classes				100
100	50	Discrete with 4 classes		19	81	
		Discrete with 5 classes		1	70	29
	10	Discrete with 4 classes	79	20	1	
		Discrete with 5 classes		1	70	29
30	50	Discrete with 4 classes	1	37	62	
		Discrete with 5 classes	5	34	58	3
	10	Discrete with 4 classes	90	9	1	
		Discrete with 5 classes	3	36	58	3

specification never overestimates the number of latent classes, which confirms that BIC is a somewhat conservative measure when deciding about the number of classes (see, for example, Dias (2004)).

As indicated earlier, in the nonparametric approach one increases the number of classes till a saturations point in reached, which seemingly lead to severely biased and much less efficient estimates. The NPML solution often consisted of a larger number of latent classes than the true discrete distribution even for the smallest level-2 sample size of 30. Such solutions contained nodes with small weights but very extreme locations, with explains the bias and inefficiency of this approach. In contrast, the semi-parametric approach will not accept such classes in the final solution because they do not yield a significantly better description of the data according to the BIC.

3.5 Conclusions

The two questions that we wished to answer based on the simulation study are 1) whether the NPML and/or semi-parametric approaches perform better in terms of bias and efficiency compared to the parametric model when the latter is misspecified, and 2) whether the NPML and/or semi-parametric approaches perform equally well in terms of bias and efficiency compared to the parametric model when the latter is correctly specified. This was studied for small and large level-1 and level-2 sample sizes and different types of random effects distributions (with a moderate ICC value). We are now able to answer these two questions for the two-level logistic regression model.

Our study showed that the NPML method gives the worst results in terms of bias and relative efficiency when compared to the parametric and semi-parametric methods, and this applies irrespective of the true random effects distribution. The semi-parametric approach performs best when the true underlying distribution of random effects is discrete. When the assumptions of the parametric model hold, the parametric approach is the best for the fixed effects estimation, but the semi-parametric approach is the preferred one for the random effects estimation. When the true distribution is exponential (continuous but not normal), the parametric model is still preferred with a small level-1 sample size, but the semi-parametric model is better with a larger level-1 sample size.

We may finally compare our conclusions with those derived from the

study by Lukočienė and Vermunt (2008) on multilevel logistic regression with only a random intercept. One important difference concerns the performance of the NPML method. Whereas this earlier study found that the NPML approach performs rather well as long as the level-1 sample size is not too small, here we have to conclude that it is by far the worst approach. In fact, the NPML method should not be used with multidimensional random effects. Another new element compared to this earlier study is that we also looked at the semi-parametric method which turned out to perform much better than the NPML method. As far as the parametric approach is concerned, similarly to the previous study it can be concluded that it is the preferred method when the normal distribution assumption holds, as well as when the distribution is continuous but not normal and the level-1 sample size is small.

One limitation of our study is that it concerned two-level regression models, and it is not clear whether our findings can be generalized to models containing more hierarchical levels. Another limitation is that we focussed on models for binary responses. The suggestion for the future research would be to look at other models from the generalized linear modeling family, as well as at models with more than two levels; that is, at the class of models described by Vermunt (2004).

In our study we investigated three different specifications for the random effects distribution: a parametric approach with an underlying normal distribution, as well as nonparametric and semi-parametric approaches using an unspecified discrete mixing distribution. As a possible alternative one may use a combination of these, namely a finite mixture

of normal distributions (Magder and Zeger, 1996; Verbeke and Molenberghs, 2000). Whereas such an approach may have particular advantages, such as that contrary to the nonparametric and semi-parametric approaches it yields nondiscrete random effects, Agresti *et al.* (2004) obtained somewhat disappointing results with this approach in the context of a log linear model for an odds ratio. Nevertheless, we believe that this hybrid approach may be promising in other situations, especially when the aim of the study is to obtain interpretable latent classes (Magidson and Vermunt, 2007) .

Chapter 4

Determining the number of components in mixture models for hierarchical data

4.1 Introduction

Vermunt (2003, 2005, 2007, 2008) proposed several types of latent class (LC) and mixture models for multilevel data sets with applications in sociological, behavioral, and medical research. Examples of two-level data sets include data from individuals (lower-level units) nested within families (higher-level units), pupils nested within schools, patients nested within primary care centers, and repeated measurements nested within individuals. A multilevel latent class model can be applied when in addition multiple responses are recorded for the lower-level units, and is thus, in fact, a model for three-level data sets. The multilevel LC models dealt with in this paper assume that lower-level units (say individuals) belong to LCs at the lower level and that higher-level units (say groups)

belong to LCs at the higher level. In other words, the models contain mixture distributions at two levels.

There is wide variety of literature available on the performance of model selection statistics for determining the number of mixture components in mixture models. The Bayesian (also known as Schwarz's) information criterion (BIC) is the most popular measure for determining the number of mixture components and it is generally considered to be a good measure (Hagenaars and McCutcheon, 2002; Nylund *et al.*, 2007). Other authors, however, prefer the Akaike information criterion (AIC) (Leroux, 1992*a*). While deciding about the number of mixture components is already a complicated task in standard mixture models, it is even more complex for multilevel mixture models. One of the difficulties consists in choosing the appropriate sample size in the BIC and CAIC formulae:

$$BIC = -2\ln L + k\ln(n) \tag{4.1}$$

and

$$CAIC = -2\ln L + k(1 + \ln(n)). \tag{4.2}$$

Here, L is the maximized value of the likelihood function for the estimated model, k is the number of free parameters to be estimated, and n is the number of observations, or equivalently, the sample size. There are several options for defining the sample size in the multilevel context, including the number of groups, the number of individuals, or either the number of groups or number of individuals depending on whether one wishes to determine the number of components at the higher or at

the lower level. Neither the literature on mixture models nor the literature on multilevel analysis give hints on what sample size to use in the computation of BIC and CAIC in multilevel mixture models.

This article presents the results of a simulation study in which we compared the performance of several methods for determining the number of mixture components in the multilevel LC models. We investigated the performance of BIC and CAIC using different sample size definitions, as well as compare BIC and CAIC with other model selection measures, such as AIC, AIC3, ICOMP (Bozdogan, 1993), and the validation log-likelihood (Smyth, 2000). Our focus is on deciding about the number of mixture components at the higher level.

The next section describes the multilevel LC model. The design of the simulation study is explained in Section 4.3. The obtained result are presented in Section 4.4. The main conclusions are highlighted in the last section.

4.2 Multilevel latent class model

Let $\boldsymbol{y}_j = (y_{j1}, \ldots, y_{ji}, \ldots, y_{jI})$ denote the vector with the I responses of individual j, $(j = 1, \ldots, n)$. A discrete LC variable is denoted by x_j, a particular LC by l_2, and the number of classes by L_2 $(l_2 = 1, \ldots, L_2)$. The basic assumptions of the LC model are: 1) that each individual belongs to (no more than) one latent class, 2) that the responses of individuals belonging to the same LC are generated by the same (probability) density, and 3) that the I responses of individual j are conditionally in-

dependent of one another given his/her class membership. Under these assumptions, the traditional LC model is defined by the following formula:

$$f(\boldsymbol{y}_j) = \sum_{l_2=1}^{L_2} P(x_j = l_2) \prod_{i=1}^{I} f(y_{ji}|x_j = l_2), \tag{4.3}$$

where $f(\boldsymbol{y}_j)$ is the marginal density of the responses of individual j, $P(x_j = l_2)$ is the unconditional probability of belonging to LC l_2, and $f(y_{ji}|x_j = l_2)$ is the conditional density for response variable i given that one belongs to LC l_2.

A multilevel LC model differs from a standard LC model in that the parameters of interest are allowed to differ randomly across groups (across higher-level units). It should be noted that the multilevel LC model is actually a model for three-level data sets; that is, for multiple responses (level-1 units) nested within individuals (level-2 units) and individuals (level-2 units) are nested within groups (level-3 units). The random variation of LC parameters across groups can be modelled using continuous or discrete group-level latent variables, or by a combination of these two. It should be noted that using the discrete latent variable approach, where parameters are allowed to differ across latent classes of groups, is similar to using a nonparametric random effects approach (Aitkin, 1999; Vermunt, 2004). In this article we focus on this discrete approach which makes use of group-level latent classes.

Let $\boldsymbol{y}_{kj} = (y_{kj1}, \ldots, y_{kji}, \ldots, y_{kjI})$ denote the I responses of individual j ($j = 1, \ldots, n_k$) from group k ($k = 1, \ldots, K$), and $\boldsymbol{y}_k = (\boldsymbol{y}_{k1}, \ldots, \boldsymbol{y}_{kj}, \ldots, \boldsymbol{y}_{kn_j})$ the full response vector of group k. The class membership of individual

j from group k is now denoted by x_{kj}. In the discrete random-effects approach it is assumed that every group belongs to one of the L_3 group-level LCs or mixture components. Let w_k denote the class membership of group k and l_3 denote a particular group-level LC ($l_3 = 1, \ldots, L_3$). The multilevel LC model can then be described by the following two equations:

$$f(\boldsymbol{y}_k) = \sum_{l_3=1}^{L_3} P(w_k = l_3) \prod_{j=1}^{n_k} f(\boldsymbol{y}_{kj}|w_k = l_3) \tag{4.4}$$

and

$$f(\boldsymbol{y}_{kj}|w_k = l_3) = \sum_{l_2=1}^{L_2} P(x_{kj} = l_2|w_k = l_3) \prod_{i=1}^{I} f(y_{kji}|x_{kj} = l_2, w_k = l_3). \tag{4.5}$$

Equation (4.4) shows how the responses of the n_k individuals belonging to group k are linked to obtain the density for the full response vector of group k, $f(\boldsymbol{y}_k)$. More precisely, it shows that the group members' responses are assumed to be mutually independent conditional on the group-level class membership. Furthermore, from Equation (4.5) it can be seen that both the lower-level mixture proportions – $P(x_{kj} = l_2|w_k = l_3)$ – and the parameters defining the response densities – $f(y_{kji}|x_{kj} = l_2, w_k = l_3)$ – may differ across higher-level mixture components.

Two interesting special cases of the multilevel LC model are obtained by constraining the terms appearing in Equation (4.5) (Vermunt, 2004, 2008). The first special case, which is the one we will use in our simulation study, is a model in which the individual-level class membership probabilities differ across group-level classes, but in which the

parameters defining the conditional distributions for the response variables do not vary across group-level classes. The latter implies that $f(y_{kji}|x_{kj} = l_2, w_k = l_3) = f(y_{kji}|x_{kj} = l_2)$. The second special case is a model in which the parameters defining the conditional distributions for the response variables differ across group-level classes, but in which individual-level class membership probabilities do not vary across group-level classes. The latter restriction implies that $P(x_{kj} = l_2|w_k = l_3) = P(x_{kj} = l_2)$. The first special case is the most natural specification if one uses the multilevel LC models a multiple-group LC model for a large number of groups. The second one is more similar to three-level random-effects regression analysis.

The unknown parameters of a multilevel LC model can be estimated by means of Maximum Likelihood (ML). For this purpose one can use the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) – the most popular algorithm for obtaining ML estimates in the context of mixture modeling – which in the context a multilevel LC model requires a specific implementation of the E step. As shown by Vermunt (2003, 2007), the relevant marginal posterior probabilities can be computed in an efficient way by making use of the conditional independence assumptions implied by the multilevel LC model. This special version of the EM algorithm, as well as a Newton-Raphson algorithm with analytic first-order derivatives and numerical second-order derivatives are implemented in the Latent GOLD software package (Vermunt and Magidson, 2008). The last version of the Latent GOLD software package was used for the realization of the simulation study reported below.

4.3 Design of the simulation study

The purpose of the simulation study was to compare the performance of different model selection indices for determining the number of mixture components at the higher-level in the multilevel LC model. These indices are BIC, AIC, AIC3, CAIC, ICOMP, and the validation log-likelihood. For BIC and CAIC we use two versions, one with the number of groups and one with the total number of individuals as the sample size.

Because we focus on detecting the correct number of group-level classes rather than on detecting the correct number of individual-level classes, we decided to keep the LC structure at the individual level fixed in our simulation design. More specifically, we used a three-class model ($L_2 = 3$) for six binary responses ($I = 6$). The class-specific "positive" response probabilities – $P(y_{kji} = 1|x_{kj} = l_2)$ – for the six items were set to $\{0.8, 0.8, 0.8, 0.8, 0.8, 0.8\}$,
$\{0.8, 0.8, 0.8, 0.2, 0.2, 0.2\}$, and $\{0.2, 0.2, 0.2, 0.2, 0.2, 0.2\}$ for LCs 1, 2, and 3, respectively. So LC 1 has a high probability of giving the positive response for all items, LC 3 a low probability for all items, and LC 2 a low probability for 3 items and a high probability for the other 3 items. Our choice of number of items, number of classes, and response probabilities is such that we obtain a condition with moderately separated classes. To give an impression of the level of the separation, our setting corresponds to an entropy based R-squared – a measure indicating how well one can predict the class memberships based on the observed responses – of about 0.63. By using moderately separated classes at the

lower level, we make sure that detection of the group-level classes is neither made too easy nor too difficult as far as this part of the model is concerned.

So far we have discussed the factors that were fixed in the simulation study. The three factors which were varied are the lower-level sample size, the higher-level sample size, and the number of LCs at the higher-level. Previous simulation studies have shown that the sample size, the number of classes, and the level of separation between the classes are the most important factors affecting the performance of model selection measures in the context mixture models (Dias, 2006). It should be noted that the separation between the higher-level classes can be manipulated in several ways; that is, by increasing the level of separation of the lower-level classes, by increasing the number of individuals per group (the lower-level sample size n_k), and by making the $P(x_{kj}|w_k)$ more different across values of w_k. We used only the lower-level sample size n_k to manipulate the level of separation. More specifically by using $n_k = 5$, 10, 15, 20 and 30 for the number of the lower-level units per higher-level unit, we created conditions ranging from very low to very high separation. The corresponding entropy-based R-squared values are given below after discussing the other design factors.

The other two factors that were varied are the higher-level sample size, for which we used $K = 50$ and 500, and the number of classes at the higher level, for which we used $L_3 = 2$ and 3. In the condition with two higher-level classes, the model probabilities were set to $P(w_k = \{1, 2\}) = \{0.5, 0.5\}$, $P(x_{kj} = \{1, 2, 3\}|w_k = 1) = \{0.2, 0.2, 0.6\}$, and $P(x_{kj} =$

$\{1,2,3\}|w_k = 2) = \{0.4, 0.4, 0.2\}$. These probabilities are such that the two LCs are moderately distinguishable. The condition with three LCs at the higher-level was created by splitting the above second class into two new classes. For this condition, the model probabilities were $P(w_k = \{1,2,3\}) = \{0.5, 0.25, 0.25\}$, $P(x_{kj} = \{1,2,3\}|w_k = 1) = \{0.2, 0.2, 0.6\}$, $P(x_{kj} = \{1,2,3\}|w_k = 2) = \{0.2, 0.6, 0.2\}$, and $P(x_{kj} = \{1,2,3\}|w_k = 3) = \{0.6, 0.2, 0.2\}$. Also here we have moderately different group-level classes. The five different n_k settings yielded entropy-based R-squared values of 0.35, 0.57, 0.71, 0.80, and 0.90 for the 2 class condition, and 0.36, 0.58, 0.73, 0.82, and 0.92 for the 3 class condition. This shows that in our settings separation was very much affected by n_k but not so much by L_3.

In total the simulation study design contained $5 \times 2 \times 2 = 20$ cells which are all possible combinations of the three design factors. For each of these cells we generate 100 data sets. With each data set we estimated multilevel LC models with a fixed number of LCs at the lower-level ($L_2 = 3$) and with different numbers of LCs at the higher-level.

4.4 Results of the simulation study

As was indicated above, the main goal of the simulation study was to determine which of the investigated model selection measures is preferable for the deciding about the number of higher-level mixture components in multilevel mixture models. For BIC and CAIC, which both have the sample size in their formula, we used two versions, one based on the num-

Table 4.1: The number of simulation replicates in which the investigated fit measure underestimated, correctly estimated, and overestimated the number of group-level mixture components for each of the three conditions.

		n_k					K		L_3		
		5	10	15	20	30	50	500	2	3	*Total*
BIC(Kn_k)	$\hat{L}_3 < L_3$	233	131	67	18	1	400	50	136	314	450
	$\hat{L}_3 = L_3$	167	269	333	382	399	600	950	864	686	1550
	$\hat{L}_3 > L_3$	0	0	0	0	0	0	0	0	0	0
BIC(K)	$\hat{L}_3 < L_3$	199	83	26	6	0	286	28	93	221	314
	$\hat{L}_3 = L_3$	201	317	374	394	399	713	972	907	778	1685
	$\hat{L}_3 > L_3$	0	0	0	0	1	1	0	0	1	1
CAIC(Kn_k)	$\hat{L}_3 < L_3$	253	146	81	33	5	456	62	153	365	518
	$\hat{L}_3 = L_3$	147	254	319	367	395	544	938	847	635	1482
	$\hat{L}_3 > L_3$	0	0	0	0	0	0	0	0	0	0
CAIC(K)	$\hat{L}_3 < L_3$	228	101	46	9	0	337	47	114	270	384
	$\hat{L}_3 = L_3$	172	299	354	391	400	663	953	886	730	1616
	$\hat{L}_3 > L_3$	0	0	0	0	0	0	0	0	0	0
AIC	$\hat{L}_3 < L_3$	103	45	9	3	0	158	5	41	122	163
	$\hat{L}_3 = L_3$	278	320	344	349	355	766	880	853	793	1646
	$\hat{L}_3 > L_3$	16	35	47	48	45	76	115	106	85	191
AIC3	$\hat{L}_3 < L_3$	155	68	13	5	0	236	5	70	171	241
	$\hat{L}_3 = L_3$	245	323	375	389	385	745	972	904	813	1717
	$\hat{L}_3 > L_3$	0	9	12	6	15	19	23	26	16	42
ICOMP	$\hat{L}_3 < L_3$	208	85	20	4	0	274	43	91	226	317
	$\hat{L}_3 = L_3$	191	310	380	392	398	714	957	900	771	1671
	$\hat{L}_3 > L_3$	1	5	0	4	2	12	0	9	3	12
Validation	$\hat{L}_3 < L_3$	78	37	9	1	0	121	4	46	79	125
log-likelihood	$\hat{L}_3 = L_3$	215	239	272	286	291	582	721	691	612	1303
	$\hat{L}_3 > L_3$	107	124	119	113	109	297	275	263	309	572

ber of higher-level observations (K) and one based on the total number of lower-level observations (Kn_k).

Table 4.1 reports the results of our simulation study per design factor aggregated over the other two design factors. For each level of the three design factors and for each investigated fit measure, we indicate the number of simulation replications in which the true number higher-level latent classes was underestimated ($\hat{L}_3 < L_3$), estimated correctly ($\hat{L}_3 = L_3$), and overestimated ($\hat{L}_3 > L_3$).

Let us first have a look at the results for BIC and CAIC using the two different definitions for the sample size. From the results in Table 4.1, one can easily see that both for BIC and CAIC using the number of groups as sample size is the best option. Underestimation of the number of mixture components it is much more likely with BIC(Kn_k) or CAIC(Kn_k) than with BIC(K) or CAIC(K). This is especially true for the conditions corresponding to low or moderate levels of separation (small or middle n_k values), as well as for the smaller higher-level sample size.

Comparison of the results of all eight investigated fit measures shows that overall AIC3 performs best. The results for BIC(K), CAIC(K), ICOMP are similar in the sense that they perform best when the number of individuals per group (the level of separation) is large enough ($n_k \geq 15$). AIC, on the other hand, performs best when separation is weak ($n_k = 5$) and when the sample size is small. As was found in other studies, AIC3 seems to provide a compromise between these two sets of measures (Dias, 2006). In contrast to our expectations, the validation

log-likelihood did not perform very well: it tends to overestimate the number of mixture components under all conditions.

4.5 Conclusions

Based on the simulation study we can draw two important conclusions. The first concerns the preferred sample size definition in the BIC and CAIC formulae. Our results show clearly that it is much better to use the number of higher-level units as the sample size instead of the total number of lower-level unit. Using the latter makes it much more likely that the number of mixture components is underestimated, especially if the separation between components is weak or moderate.

The second set of conclusions concern the comparison of all investigated measures. These results are very much in agreement with what is known from simulation studies on standard mixture models. BIC, CAIC, and ICOMP perform very well when the level of separation and the sample size are large enough. In contrast, AIC seems to be the preferable method when the sample size is small and when the level of separation is low. AIC3 offers a good compromise between the tendency of BIC, CAIC, and ICOMP to underestimate the number of mixture components with low separation and small sample sizes and the tendency of AIC to underestimate the number of mixture components with higher separation and large sample sizes.

As in any simulation study, we had to make various choices which limit the range of our conclusions. First of all, we concentrated on selecting

the number of classes at the higher level assuming that the number of classes at the lower level is known. Further research is needed to determine whether the same conclusions apply for selecting the number of lower-level classes, or for selecting simultaneously the number of lower- and higher-level classes. Second, we used a classical LC model for binary responses whereas multilevel mixture models can also be used with other types of response variables. Finally, we concentrated on the variant of the multilevel LC model in which only the lower-level class proportions differ across higher-level classes. As was shown when introducing the model, other multilevel LC models may assume that response variables are directly related to the group-level class membership. It seems to be useful to replicate our simulation study for other types of multilevel mixture models, as well as for response variables of other scale types.

Chapter 5

The simultaneous decision about the number of lower- and higher-level classes in multilevel latent class analysis

5.1 Introduction

During the last decades, latent class (LC) analysis has become part of the standard statistical toolbox of researchers in applied areas such as medicine, biology, social sciences, psychology, education, criminology, and marketing. As is typical for most statistical techniques, one of the assumptions in LC modeling is that the available sample consists of a set of independent units, an assumption which is inadequate when units are nested within clusters sharing common environments, experiences, and interactions. In such situations, one should use multilevel techniques which take the dependencies between lower-level units resulting from the hierarchical data structure into account (Hox, 2002; Snijders and Bosker,

1999).

Recently, various types of multilevel extensions of LC and other types of finite mixture models have been developed (Asparouhov and Muthén, 2008; Di and Bandeen-Roche, 2008; Palardy and Vermunt, in press; Vermunt, 2003, 2004, 2007, 2008). The common element of these extensions is that some of the LC model parameters are allowed to vary randomly across higher-level units. Although group differences could also be modelled using multi-group LC analysis (Clogg and Goodman, 1984), such an approach is only feasible when the number of groups is not too large, say between two and ten. With larger numbers of (possibly small) groups, it is more appropriate to model group differences using random effects. The multilevel LC model proposed by Vermunt (2003) involves expanding the standard LC model with either a continuous or a discrete latent variable at the higher level, yielding either a parametric or a non-parametric specification for the random effects distribution (Aitkin, 1999; Skrondal and Rabe-Hesketh, 2004).

This paper deals with the non-parametric (or semi-parametric) variant of the multilevel LC model in which differences across groups are modelled using a discrete latent variable at the group level. Applications of this variant of the multilevel LC model typically aim at simultaneously clustering individuals and groups; that is, lower-level units are assumed to belong to lower-level LCs differing in the distribution of the observed responses and higher-level units are assumed to belong to higher-level LCs differing in the distribution of the lower-level LCs. A good example is a recent study by Cavrinia *et al.* (2009) on patients' satisfaction

with hospital services: the lower-level LCs represent clusters of patients with similar satisfaction levels concerning the studied aspects of hospital services, and LCs at the higher level represent clusters of hospitals with similar distributions of patients across the patient-level satisfaction clusters. Other applications of this variant of the multilevel LC model include studies by Bassi (2009); Bijmolt *et al.* (2004); Bouwmeester *et al.* (2007); Kragelj and Schlutter (2007); Pirani *et al.* (2009); Rindskopf (2006).

Even though the theory of multilevel LC analysis is well developed and interesting applications have already been published in a broad range of applied areas, one important issue has received little attention, namely, the problem related to the simultaneous decision regarding the number of lower- and higher-level LCs. For standard LC and mixture models, there is a large body of literature on the performance of statistics for determining the number of mixture components. It is well-known that asymptotic likelihood ratio tests can not be used because certain regularity conditions do not hold, but that approximate p-values can be obtained using bootstrap procedures (McLachlan, 1987; McLachlan and Peel, 2000). However, because bootstrapping is computationally very intensive, applied researchers typically prefer using measures weighting model fit (the log-likelihood value) and model complexity (the number of parameters). The most popular of these measures is the Bayesian information criterion (BIC) (Hagenaars and McCutcheon, 2002; Nylund *et al.*, 2007; Schwarz, 1978). Other authors, however, suggest using the Akaike information criterion (AIC) (Akaike, 1974), at least in particu-

lar situations (Lin and Dayton, 1997). Other alternatives are adjusted versions of AIC, such as consistent AIC (CAIC) (Bozdogan, 1987) and AIC3 (Bozdogan, 1993).

While deciding about the number of mixture components is already a rather complex task in standard LC and mixture modeling, it is even more complex in multilevel mixture modeling. It not only involves two instead of one decision, about both the number of lower- and higher-level LCs, these decisions may also be mutually dependent. Except for the simulation study by Lukočienė and Vermunt (2009), this issue has not received any attention in the literature on multilevel LC analysis. However, these authors focussed on the rather simplified situation in which the number of lower-level classes is known; that is, on the situation in which only one decision (about the higher-level classes) has to be made. Their simulation study showed that overall AIC3 performs best. Another important result is that the sample size in the BIC and CAIC formulae should be the number of higher-level units.

The aim of the current article is to extend the work of Lukočienė and Vermunt (2009) to the situation we encounter in practice in which both the number of higher- and lower-level LCs is unknown. In other words, we focus on the simultaneous decision about the number of mixture components at the lower and higher level. We compare the performance of the most popular measures: BIC, AIC, CAIC and AIC3, we investigate the performance of BIC and CAIC using different sample size definitions, and we propose a stepwise model fitting strategy that makes the application of the multilevel LC model easier. The theory is illustrated

using an application on the job satisfaction of University of Florence graduates from different degree programs, where the aim is to cluster both graduates and programs into homogeneous LCs.

The next section describes the multilevel LC model. The new three-step model fitting procedure and the model selection criteria that will be evaluated are described in Section 5.3. Sections 5.4 and 5.5 present the design and the results of the simulation study. The application is presented in Section 5.6. The last section contains the main conclusions of our study.

5.2 The multilevel latent class model

We denote the observed responses in a data set used to build a multilevel LC model by y_{kji}, where the indices i, j, and k refer to a response variable, an individual or lower-level unit, and a group or higher-level unit, respectively. The number of response variables equals I ($i = 1, \ldots, I$), the number individuals within group k equals n_k ($j = 1, \ldots, n_k$), and the number of groups equals K ($k = 1, \ldots, K$). Moreover, the total number of lower-level units equals $N = \sum_{k=1}^{K} n_k$. The vectors $\boldsymbol{y}_{kj} = (y_{kj1}, \ldots, y_{kji}, \ldots, y_{kjI})$ and $\boldsymbol{y}_k = (\boldsymbol{y}_{k1}, \ldots, \boldsymbol{y}_{kj}, \ldots, \boldsymbol{y}_{kn_k})$ contain the I responses of individual j from group k and the full set responses of group k, respectively. Note that such a data set can be perceived as either a I-variate two-level data set or a univariate three-level data set.

A multilevel LC model assumes that individuals belong to one of L LCs and groups to one of H LCs. The variables representing the lower-

and higher-level class memberships are denoted by x_{kj} and w_k, respectively, and a particular class by l $(l = 1, \ldots, L)$ and h $(h = 1, \ldots, H)$, respectively.

The multilevel LC model proposed by Vermunt (2003, 2008) can be formulated using two basic equations. The first equation defines the (mixture) model for $f(\boldsymbol{y}_k)$, the marginal density of the full response vector of group k; that is,

$$f(\boldsymbol{y}_k) = \sum_{h=1}^{H} P(w_k = h) \prod_{j=1}^{n_k} f(\boldsymbol{y}_{kj} | w_k = h). \tag{5.1}$$

Here, $P(w_k = h)$ is the probability that group k belongs to LC h and $f(\boldsymbol{y}_{kj}|w_k = h)$ is the conditional density for the response vector of individual j in group k conditional on the membership of group k to LC h. The second equation defines the (mixture) model for $f(\boldsymbol{y}_{kj}|w_k = h)$; that is,

$$f(\boldsymbol{y}_{kj}|w_k = h) = \sum_{l=1}^{L} P(x_{kj} = l|w_k = h) \prod_{i=1}^{I} f(y_{kji}|x_{kj} = l, w_k = h), \tag{5.2}$$

where $P(x_{kj} = l|w_k = h)$ is the probability that individual j of group k belongs to LC l given that the group belongs to LC h, and $f(y_{kji}|x_{kj} = l, w_k = h)$ is the conditional density for response variable i of individual j in group k given the membership to individual-level class l and group-level class h.

These two equations clearly show which conditional independence assumptions are made in a multilevel LC analysis. First, the observations of the n_k individuals in group k are assumed to be independent of one

another given the group-level class membership. Note that this assumption is typical for any type of multilevel analysis: observations are assumed to be independent conditional on the random effects (Skrondal and Rabe-Hesketh, 2004). Second, the I responses of individual j are assumed to be independent of each other given the group and individual LC memberships, which is the basic assumption of most LC models and usually referred to as the local independence assumption (Bartholomew and Knott, 1999; Hagenaars and McCutcheon, 2002).

The last element in the specification of a multilevel LC model is the specification of the conditional densities $f(y_{kji}|x_{kj} = l, w_k = h)$, which will typically be assumed to belong to the exponential family. This can, for example, be a normal or gamma distribution for continuous responses, a Poisson, binomial, or negative binomial distribution for counts, and a multinomial distribution for categorical responses. In the current paper, we restrict ourselves to models for categorical responses, which means that $y_{kji} = 1, \ldots, M_i$, where M_i is the number of categories of the ith response variable. The multinomial form of density $f(y_{kji}|x_{kj} = l, w_k = h)$ can be expressed as

$$f(y_{kji}|x_{kj} = l, w_k = h) = \prod_{m=1}^{M_i} (\pi_{imlh})^{d_{kjim}}, \tag{5.3}$$

where d_{kjim} represent an indicator variable taking on the value 1 if $y_{kji} = m$ and 0 otherwise, and where π_{lhim} represents a multinomial probability subject to the constraints $\pi_{lhim} \geq 0$ and $\sum_{m=1}^{M_i} \pi_{lhim} = 1$.

Equations (5.1) and (5.2) describe the multilevel LC model in its

most general form; that is, as a model in which both the lower-level mixture proportions – $P(x_{kj} = l|w_k = h)$ – and the parameters defining the response densities – $f(y_{kji}|x_{kj} = l, w_k = h)$ – are allowed to differ across higher-level classes. Most applications of multilevel LC analysis, however, use one of two more restricted special cases. More specifically, they impose one of the following two constraints:

(1) $f(y_{kji}|x_{kj} = l, w_k = h) = f(y_{kji}|x_{kj} = l)$;

(2) $P(x_{kj} = l|w_k = h) = P(x_{kj} = l)$.

In the first restricted special case, $P(x_{kj} = l|w_k = h)$ is estimated freely, but the parameters defining the conditional distributions are assumed to be independent of the higher-level class membership (Vermunt, 2003, 2008). This structure is the one used in all the applications listed in the introduction, that is, in applications aiming at the simultaneous clustering of higher- and lower-level units. In fact, the clustering of higher-level units is performed by "pushing up" the information contained in the multiple lower-level responses via the lower-level class memberships. In the second special case, the parameters defining $f(y_{kji}|x_{kj} = l, w_k = h)$ are estimated freely, but the lower-level class membership is assumed to be independent of the higher-level class membership (Vermunt, 2004). This specification is in fact similar to the variance decomposition used in the three-level regression models: the variation in the responses is split into a between-group part and a within-group part (Skrondal and Rabe-Hesketh, 2004). In our simulation study, we focus on the first specification, which has proven to be the most useful one in applied research

(see Section 5.1).

Before discussing in more detail the issue of model selection, we would like to stress that deciding about the number of mixture components is not always an issue in (multilevel) LC or mixture modeling. It is, of course, an issue when the model is used as a cluster technique with the aim to find a good fitting and easy to interpret solution. However, mixture models can also be used as random effect models with a non parametric specification of the random-effects distribution (Aitkin, 1999). In such applications, one should increase the number of LCs until the log-likelihood function reaches its maximum.

5.3 Determining the number of lower- and higher-level classes

5.3.1 A three-step model fitting procedure

Determining the number of classes in multilevel LC analysis involves a simultaneous decision regarding the number of LCs at multiple levels of the hierarchical structure. The main complicating factor is that these decisions may be mutually dependent.

The model fitting strategy used in the first paper on multilevel LC analysis (Vermunt, 2003) – and which is also the strategy used in most applications of this model – is, in fact, a two-step procedure. One first determines the number of lower-level classes ignoring the multilevel structure, and subsequently determines the number of higher-level classes fixing the number of lower-level classes at the value from the first step.

It should be noted that the simulation study by Lukočienė and Vermunt (2009) on the selection of the number of higher-level classes builds on this model selection strategy in that it investigates the performance of various model selection criteria in the second step. The main disadvantage of this two-step strategy is that it accounts only partially for the dependency between the two decisions to be made. More specifically, the dependency of the decision about the number of lower-level classes on the selected number of higher-level classes is fully ignored.

Bijmolt *et al.* (2004) used an alternative model fitting strategy which involves estimating the multilevel LC model for all relevant combinations of L and H. In their application, this implied estimating models with L ranging from 1 to 15 and H ranging from 1 to 8. Vermunt (2008) used the same procedure in a set of applications illustrating the use of multilevel LC analysis in medical research. The two main disadvantages of this procedure are that it may require estimating a large number of models (more than 100 in the Bijmolt *et al.* (2004) application) and that it does not allow using different measures when deciding about the value of L and H.

We propose an alternative three-step model fitting procedure which 1) is less computationally intensive than the procedure by Bijmolt *et al.* (2004), 2) accounts for the fact that the value of L may depend on the selected value of H, and 3) allows using different measures when deciding about L and H. This procedure consists of the following three steps:

(1) determine the number of lower-level classes ignoring the multilevel structure (that is, assuming that $H = 1$);

(2) fix the number of lower-level classes to the value of step 1 and determine the number of higher-level classes;

(3) fix the number of higher-level classes to the value of step 2 and redetermine the number of lower-level classes.

Note that the first two steps are the same as the ones used by Vermunt (2003), but with the important modification that different fit indices may be used in steps 1 and 2 (more details are provided below). The aim of the extra step 3 is to evaluate whether the number of lower-level classes changes after taking into account the dependencies between lower-level units due to the multilevel data structure. Of course, a fourth step could be added in which the number of higher-level classes is reevaluated fixing L to the value of step 3, as well as a fifth step in which the number of lower-level classes is reevaluated fixing H to the value of step 4, etc.. In the current study, we, however, restrict ourselves to the above three-step approach, which we believe already provides an important improvement compared to the Bijmolt *et al.* (2004); Vermunt (2003) approaches.

5.3.2 Model selection measures

When working within a maximum likelihood estimation framework as we do here, comparison of nested models is typically performed by means of likelihood-ratio tests which under certain regularity conditions follow a chi-squared distribution. However, a problem is that such likelihood-ratio tests can not be used for comparing models with different numbers of classes because the null model with the smaller number of classes is

obtained by fixing one or more parameters of the alternative model at their boundary values. A solution proposed by various authors is to approximate the p-value associated with these likelihood-ratio tests using parametric bootstrap procedures (see, for example, McLachlan (1987); Nylund *et al.* (2007)). However, these bootstrap-based testing procedures are seldom used by applied researchers because they are computationally very intensive and, moreover, their correct implementation is not at all straightforward.

Most researchers applying LC analysis will make use of information criteria, which are measures weighting model fit (the log-likelihood value) and model complexity (the number of parameters). As the log-likelihood will typically increase with increasing model complexity (with increasing number of classes), it is penalized by the addition of a term measuring the complexity of the model. These information criteria can be expressed most generally as follows:

$$IC = -2 \ln L(\boldsymbol{\theta}) + Cr,$$

where $L(\boldsymbol{\theta})$ is the maximized log-likelihood value for a model with parameters $\boldsymbol{\theta}$, r is the number of independent parameters in this model, and C is the weight given to the penalty term based on r. The lower is the value of an information criterion, the better the model. The various information criteria proposed in the literature differ in the value of C.

Most texts on LC analysis suggest using the BIC for deciding about the number of classes (see, for example, Hagenaars and McCutcheon

(2002); Magidson and Vermunt (2004)). BIC is defined as follows:

$$BIC = -2 \ln L(\boldsymbol{\theta}) + \ln(n)r \tag{5.4}$$

where n is the number of observations (sample size). Simulation studies have shown that usually BIC performs very well, but also that it may sometimes underestimate the number of classes, namely, when classes are not well separated (see, for example, Dias (2004); Nylund *et al.* (2007)).

Others suggest using the AIC (Akaike, 1974), which is expressed as

$$AIC = -2 \ln L(\boldsymbol{\theta}) + 2r. \tag{5.5}$$

Simulation studies have shown that AIC tends to overestimate the number of classes (Dias, 2004; McLachlan and Peel, 2000), although others report that AIC works well in specific situations (Lin and Dayton, 1997).

Bozdogan proposed two adjusted versions of AIC: AIC3 (Bozdogan, 1993) and CAIC (Bozdogan, 1987), which are used more and more in LC analysis. AIC3 and CAIC can be expressed, respectively, by:

$$AIC3 = -2 \ln L(\boldsymbol{\theta}) + 3r \tag{5.6}$$

and

$$CAIC = -2 \ln L(\boldsymbol{\theta}) + (1 + \ln(n))r. \tag{5.7}$$

Simulation studies by Andrews and Currim (2003); Dias (2004) showed that AIC3 is the best performing criterion in LC analysis with categorical response variables. Note that the AIC3 weight of 3 is typically in between

the BIC weight of $\ln n$ and the AIC weight of 2. It can thus be seen as a compromise between these two measures that, compared to BIC, is better able to detect badly separated classes and that, contrary to AIC, is less likely to come up with spurious classes. The reported behavior of CAIC is similar to the one of BIC, which is not surprising given that their penalties are rather similar.

Lukočienė and Vermunt (2009) pointed at a specific issue when using BIC and CAIC in the context of multilevel analysis: it is not clear whether the sample size should be the number of groups (K), the total number of individuals (N), or either the number of groups or number of individuals depending on whether one wishes to test model features related to the higher or lower level.

The aim of the current study is to determine the performance of the various information criteria described above for deciding about the number of classes in multilevel LC models. The work by Lukočienė and Vermunt (2009) is the only study that has been published on this topic. This study, however, restricted itself to the simplified situation in which the number of lower-level classes can be assumed to be known. The results of this study can be assumed to be valid in step 2 of the three-step model fitting procedure described above, but only if L was correctly estimated in step 1. The two main results of the Lukočienė and Vermunt (2009) study are 1) that K should be used as the sample size in the BIC and CAIC formulae when deciding about the number of higher-level classes, and 2) that overall, as in standard LC models, AIC3 is the preferred measure.

The current study aims at providing information on the performance of the various information criteria in the more realistic situation in which the number lower-level latent classes is unknown. We will again address the issue related to sample size definition in BIC and CAIC, but now not only for the selection of the number of higher-level classes, but also for the selection of the number of lower-level classes. Moreover, we will investigate the possibility of using different sample size definitions in steps 1 and 3 on the one hand and step 2 on the other hand.

5.4 Design of the simulation study

The two main questions we would like to address in our simulation study are:

- How well does the proposed three-step model fitting procedure perform under the studied conditions?
- How well do the various information criteria perform under the studied conditions?

With performance we mean whether the model with the correct number of LCs is selected by our procedure. The starting point for the design of the simulation study – for the definition of the conditions that will be varied in our study – is what is known from previous simulation studies on determining the number of LCs in latent class models. As summarized by Dias (2004), the three main factors determining the difficulty of detecting the correct number of classes are:

(1) the number of classes (the larger the number of classes the less likely that one finds the right number of classes),

(2) the separation between the classes (the smaller the separation between the classes the less likely that one finds the right number of classes).

(3) the sample size (the smaller the sample size the less likely that one finds the right number of classes),

These are the key factors that will be manipulated, and because we are dealing with a multilevel LC model instead of a standard LC model, these will be manipulated for both the higher- and lower-level classes.

It should be noted that while "separation between the classes" has been reported to be the most important factor (see, Andrews and Currim (2003); Dias (2004); Sarstedt (2008)), it is also a somewhat "obscure" factor because it can be manipulated and quantified in various ways. As is often done in LC and mixture modeling, we will quantify the separation between classes using an entropy based R-squared measure indicating how well the class membership can be predicted from the observed responses (Wedel and Kamakura., 1998). A value of 0 corresponds to a prediction that is no better than change (and thus no separation at all) and a value of 1 to a perfect prediction (and thus a perfect separation). Below we provide more details on how the separation between higher- and lower-level classes is manipulated.

Before describing how the three relevant factors mentioned above were varied, we would like to mention what was kept fixed within our simula-

Table 5.1: Assumed probability of belonging to a particular higher-level class [$P(w_k = h)$] for the $H = 2$ (a) and $H = 3$ (b) conditions.

h	
1	2
0.5	0.5

(a)

h		
1	2	3
0.5	0.25	0.25

(b)

tion study. What is kept fixed is the number of response variables and the number of categories of the response variables. More specifically, we used a LC model with six binary responses ($I = 6$ and $M_i = 2$). The reason for keeping these two factors fixed is that these are factors mainly affecting the separation between the lower-level classes; that is, the larger the I and M_i values, the larger the separation between the classes. However, as explained below, the lower-level class separation can be manipulated in a much simpler and direct way.

The factor number of classes is the most easily manipulated. More specifically, the number of LCs were varied from two to three at both levels ($L = 2, 3$ and $H = 2, 3$). Table 5.1 reports the assumed values for the probability of belonging to a particular higher-level class [$P(w_k = h)$] under the $H = 2$ and $H = 3$ conditions. Below, we will discuss the settings for $P(x_{kj} = l|w_k = h)$, which is one of the factors affecting the higher-level separation.

The separation between the lower-level classes is most easily manipulated via the class-specific response probabilities. Table 5.2 presents the overall structures used for the class-specific probability of responding in the second (say the positive) category – referred to as π_{i2l} in Equation (5.3) – in the $L = 2$ and $L = 3$ conditions. Note that $\pi_{i1l} = 1 - \pi_{i2l}$.

Table 5.2: Assumed structure for the class-specific "positive" response probabilities (π_{i2l}) for the $L = 2$ (a) and $L = 3$ (b) conditions.

	l	
i	1	2
1	p	$1-p$
2	p	$1-p$
3	p	$1-p$
4	p	$1-p$
5	p	$1-p$
6	p	$1-p$

(a)

	l		
i	1	2	3
1	p	p	$1-p$
2	p	p	$1-p$
3	p	p	$1-p$
4	p	$1-p$	$1-p$
5	p	$1-p$	$1-p$
6	p	$1-p$	$1-p$

(b)

These structures are such that only one parameter (denoted by p) needs to be specified. More specifically, the settings $p = 0.8$ and $p = 0.9$ define the two conditions for the lower-level separation. Irrespective of the setting for p, the interpretation of the classes is as follows: the first LC has the higher probability of giving a positive response for all items, the last LC has the lower probability of giving a positive response for all items, and, in the $L = 3$ condition, the middle LC has the higher probability for 3 items and the lower probability for the other 3 items. The entropy based R-squared value is around 0.63 for the $p = 0.8$ condition and around 0.88 for the $p = 0.9$ condition, which applies both for the $L = 2$ and $L = 3$ conditions. This is the range of separation levels that we typically encounter in LC analysis applications.

A complicating factor in the setting of the separation values is that these are not independent across hierarchical levels. Because lower-level observations are correlated, the actual lower-level separation values will be somewhat larger than what we reported above; how much larger depends on the separation of the higher-level classes. The reversed dependency also applies, increasing the lower-level separation also increases

Table 5.3: Assumed values for the lower-level LC probabilities conditional the higher-level class – ($P(x_{kj} = l|w_k = h)$) – for $L = 2, 3$ and $H = 2, 3$.

	h	
l	1	2
1	0.3	0.6
2	0.7	0.4

(a)

	h		
l	1	2	3
1	0.3	0.5	0.7
2	0.7	0.5	0.3

(b)

	h	
l	1	2
1	0.3	0.6
2	0.3	0.2
3	0.4	0.2

(c)

	h		
l	1	2	3
1	0.3	0.5	0.7
2	0.3	0.3	0.1
3	0.4	0.2	0.2

(d)

the higher-level separation.

The separation between the higher-level classes can be manipulated in various ways; that is, by increasing the level of separation of the lower-level classes, by making the $P(x_{kj} = l|w_k = h)$ more different across values of h, and by increasing the number of individuals per group (the lower-level sample size n_k). For the lower-level separation, we already presented the two settings ($p = 0.8$ and $p = 0.9$). Table 5.3 presents the settings for the conditional probabilities $P(x_{kj} = l|w_k = h)$. From the information in Tables 5.1 and 5.3, it can be seen that the $H = 3$ conditions were created by splitting the second class of the $H = 2$ conditions into two (equal size) classes, and the $L = 3$ conditions by splitting the second class of the $L = 2$ conditions into two classes.

The main factor that we used to manipulate the higher-level separation is the number of lower-level units per group (the lower-level sample size n_k). Note that the larger the number of units per group, the more information we have about the group-level class membership. More specifically, we used the values $n_k = 5$, 10, 15, 20, and 30 to create

Table 5.4: Entropy-based R-squared values for the higher-level classes for all combinations of p, H, L, and n_k.

			n_k				
p	H	L	5	10	15	20	30
0.9	2	2	0.27	0.46	0.60	0.69	0.82
		3	0.21	0.35	0.45	0.53	0.64
	3	2	0.27	0.45	0.59	0.69	0.82
		3	0.23	0.40	0.51	0.60	0.73
0.8	2	2	0.24	0.42	0.55	0.65	0.78
		3	0.19	0.32	0.42	0.49	0.60
	3	2	0.22	0.39	0.51	0.61	0.75
		3	0.18	0.32	0.43	0.51	0.63

conditions ranging from very low to moderately high separation. The entropy-based R-squared values reported in Table 5.4 show that higher-level class separation is affected strongly by the value n_k, somewhat by the values of p and L, and very weakly by the value of H.

The last factor that was varied is the higher-level sample size, for which we used $K = 30$, 100, and 1000. These sample sizes were chosen to cover the full range of small, moderate, and large sample sizes encountered in multilevel applications in biomedical, behavioral, and social science research.

In total, the simulation study design contained $2 \times 2 \times 2 \times 5 \times 3 = 120$ cells representing all possible combinations of the five varied design factors. For each of these cells we generated 10 data sets. The syntax version of the Latent GOLD (Vermunt and Magidson, 2008) was used for the realization of the simulation study, as well as for the real data analysis reported below.

5.5 Results of the simulation study

The results of the simulation study will be presented in three stages. First, we discuss the results for the lower-level classes; that is, of steps 1 and 3 of our three-step procedure. Then, we look at the higher-level classes (step 2 of our procedure). We end with the description of the overall results concerning the simultaneous decision about L and H.

5.5.1 Results for lower-level classes

Table 5.5 presents the results of the simulation study for the lower-level obtained after step 3. Per design factor and information criterion, it reports the number of simulation replications in which the number of classes is underestimated ($\hat{L} < L$), correctly estimated ($\hat{L} = L$), and overestimated ($\hat{L} > L$). Before looking at these results we would like to repeat that the main factors that we intended to manipulate are sample size, number of classes, and separation between classes. The total sample size at the lower-level (N) is affected by both n_k and K, while L and p represent the number of classes and separation conditions. However, because increasing n_k increases the higher-level separation, it may also increase somewhat the lower-level separation. The factor H is less relevant for the lower-level results.

The lower-level results are very much in agreement with simulation results for standard latent class models (Andrews and Currim, 2003; Dias, 2004; Sarstedt, 2008). Indeed, sample size, number of classes, and separation between classes affect the difficulty of finding the correct

Table 5.5: Number of simulation replications in which the number of lower-level classes is underestimated, correctly estimated, and overestimated.

		n_k					K			H		L		p	
		5	10	15	20	30	30	100	1000	2	3	2	3	0.8	0.9
BIC(K)	$\hat{L} < L$	11	0	0	0	0	11	0	0	6	5	0	11	11	0
	$\hat{L} = L$	229	240	239	240	240	388	400	400	593	595	599	589	588	600
	$\hat{L} > L$	0	0	1	0	0	1	0	0	1	0	1	0	1	0
BIC(N)	$\hat{L} < L$	18	7	1	0	0	25	1	0	13	13	0	26	26	0
	$\hat{L} = L$	222	233	239	240	240	375	399	400	587	587	600	574	574	600
	$\hat{L} > L$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CAIC(K)	$\hat{L} < L$	15	4	0	0	0	19	0	0	11	8	0	19	19	0
	$\hat{L} = L$	225	236	240	240	240	381	400	400	589	592	600	581	581	600
	$\hat{L} > L$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CAIC(N)	$\hat{L} < L$	19	16	2	0	0	36	1	0	17	20	0	37	37	0
	$\hat{L} = L$	221	224	238	240	240	364	399	400	583	580	600	563	563	600
	$\hat{L} > L$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AIC	$\hat{L} < L$	1	0	1	2	2	1	1	4	0	6	0	6	1	5
	$\hat{L} = L$	208	215	216	208	207	338	352	364	525	529	525	529	514	540
	$\hat{L} > L$	31	25	23	30	31	61	47	32	75	65	75	65	85	55
AIC3	$\hat{L} < L$	10	0	0	1	0	10	0	1	5	6	0	11	10	1
	$\hat{L} = L$	228	237	237	237	237	385	395	396	590	586	593	583	580	596
	$\hat{L} > L$	2	3	3	2	3	5	5	3	5	8	7	6	10	3

model and, moreover, AIC3 is the best performing criterion. BIC(N) and CAIC(N) are more likely than AIC3 to underestimate the number of classes with smaller samples, larger number of classes, and lower separation levels. Moreover, AIC is more likely to overestimate the number of classes in all situation, but more strongly with smaller samples sizes, smaller number of classes, and smaller separation.

Comparison of the performance of the somewhat unconventional BIC(K) and CAIC(K) measures with the AIC3, BIC(N), and CAIC(N) shows that these perform almost as well as AIC3, and thus better than BIC(N), and CAIC(N). This somewhat surprising result can easily be explained: under all conditions the weight $\ln K$ is closer to the AIC3 weight of 3 than $\ln N$.

Whereas Table 5.5 provides the results obtained after step 3, we also looked at the results after step 1. We, however, did not encounter significant differences between step 1 and step 3 in the aggregated results. The reason for this is that step 3 can be expected to have the most impact when the lower-level model is weak (low separation between classes and small total sample size $K \cdot n_k$) and the higher-level model strong (high separation between group-level classes). However, in the investigated settings, lower-level classes were never badly separated, and in the larger n_k conditions (when the higher-level separation is large) the total lower-level sample size was never very small. Nevertheless, comparison of the $K = 30$ and $n_k = 30$ with the $K = 100$ and $n_k = 10$ conditions – which are similar in terms of total lower-level sample size but differ in terms of higher-level separation – shows that step 3 may yield some

improvement compared to step 1, at least for the "weaker" measures AIC, BIC(N), and CAIC(N).

5.5.2 Results for higher-level classes

Table 5.6 presents the results of the simulation study for the higher-level obtained after step 2. Per design factor and information criterion, it reports the number of simulation replications in which the number of classes is underestimated ($\hat{H} < H$), estimated correctly ($\hat{H} = H$), and overestimated ($\hat{H} > H$). As mentioned earlier, the key factors expected to affect the performance of the various information criteria are sample size, number of classes, and separation between classes. For the higher level, sample size is K, number of classes is H, and separation depends most strongly on n_k and somewhat on L, and p. The results of Table 5.6 show that each of the investigated criteria perform better under the easier conditions (larger sample, fewer classes, and larger separation between classes). Another thing that can be observed is that much lower percentages of correctly estimated numbers of LCs are obtained than for the lower-level part of the model. The explanation for this is that overall the separation values we used for the higher level were much lower than the ones for the lower level, which means that we are on average dealing with more difficult situations.

Comparing the various information criteria with one another shows that overall AIC and AIC3 perform better than the other criteria. Moreover, AIC performs better than AIC3 in the situations in which it is more difficult to demonstrate the existence of H classes (smaller n_k, smaller

Table 5.6: Number of simulation replications in which the number of higher-level classes is underestimated, correctly estimated, and overestimated.

		n_k					K			H		L		p	
		5	10	15	20	30	30	100	1000	2	3	2	3	0.8	0.9
BIC(K)	$\hat{H} < H$	192	138	110	96	58	261	223	110	115	479	311	283	316	278
	$\hat{H} = H$	48	101	130	144	182	138	177	290	484	121	289	316	283	322
	$\hat{H} > H$	0	1	0	0	0	1	0	0	1	0	0	1	1	0
BIC(N)	$\hat{H} < H$	197	164	129	114	78	310	246	126	161	521	342	340	354	328
	$\hat{H} = H$	43	76	111	126	162	90	154	274	439	79	258	260	246	272
	$\hat{H} > H$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CAIC(K)	$\hat{H} < H$	196	151	114	102	66	280	236	113	134	495	324	305	332	297
	$\hat{H} = H$	44	88	126	138	174	119	164	287	465	105	276	294	267	303
	$\hat{H} > H$	0	1	0	0	0	1	0	0	1	0	0	1	1	0
CAIC(N)	$\hat{H} < H$	198	168	135	119	81	325	249	127	178	523	352	349	365	336
	$\hat{H} = H$	42	72	105	121	159	75	151	273	422	77	248	251	235	264
	$\hat{H} > H$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AIC	$\hat{H} < H$	151	101	84	65	41	220	162	60	61	381	252	190	244	198
	$\hat{H} = H$	86	130	148	163	185	170	223	319	500	212	333	379	333	379
	$\hat{H} > H$	3	9	8	12	14	10	15	21	39	7	15	31	23	23
AIC3	$\hat{H} < H$	173	116	96	73	48	248	181	77	89	417	274	232	277	229
	$\hat{H} = H$	67	122	143	165	190	150	216	321	505	182	325	362	320	367
	$\hat{H} > H$	0	2	1	2	2	2	3	2	6	1	1	6	3	4

Table 5.7: Number of simulation replications in which the number class at both levels classes is correctly estimated and not correctly estimated.

		n_k					K			H		L		p	
		5	10	15	20	30	30	100	1000	2	3	2	3	0.8	0.9
BIC(K,K)	$\hat{H}=H$ and $\hat{L}=L$	277	341	369	384	422	526	577	690	1077	716	888	905	871	922
	$\hat{H}\neq H$ and/or $\hat{L}\neq L$	203	139	110	97	58	273	224	110	122	485	311	296	329	278
BIC(N,N)	$\hat{H}=H$ and $\hat{L}=L$	265	309	350	366	402	465	553	674	1026	666	858	834	820	872
	$\hat{H}\neq H$ and/or $\hat{L}\neq L$	215	171	130	114	78	335	247	126	174	534	342	366	380	328
BIC(K,N)	$\hat{H}=H$ and $\hat{L}=L$	270	334	369	384	42	513	576	690	1071	708	889	890	857	922
	$\hat{H}\neq H$ and/or $\hat{L}\neq L$	210	146	111	96	58	287	224	110	129	492	311	310	343	278
CAIC(K,K)	$\hat{H}=H$ and $\hat{L}=L$	269	324	366	378	414	500	564	687	1054	697	876	875	848	903
	$\hat{H}\neq H$ and/or $\hat{L}\neq L$	211	156	114	102	66	300	236	113	146	503	324	325	352	297
CAIC(N,N)	$\hat{H}=H$ and $\hat{L}=L$	263	296	343	361	399	439	550	673	1005	657	848	814	798	864
	$\hat{H}\neq H$ and/or $\hat{L}\neq L$	217	184	137	119	81	361	250	127	195	543	352	386	402	336
CAIC(K,N)	$\hat{H}=H$ and $\hat{L}=L$	265	312	364	378	414	483	563	687	1048	685	876	857	830	903
	$\hat{H}\neq H$ and/or $\hat{L}\neq L$	215	168	116	102	66	317	237	113	152	515	324	343	370	297
AIC	$\hat{H}=H$ and $\hat{L}=L$	294	345	364	371	392	508	575	683	1025	741	858	908	847	919
	$\hat{H}\neq H$ and/or $\hat{L}\neq L$	186	135	116	109	88	292	225	117	175	459	342	292	353	281
AIC3	$\hat{H}=H$ and $\hat{L}=L$	295	359	380	402	427	535	611	717	1095	768	918	945	900	963
	$\hat{H}\neq H$ and/or $\hat{L}\neq L$	185	121	100	78	53	265	189	83	105	432	282	255	300	237

p, smaller K, and larger H), whereas AIC3 performs better in the easier situations, where AIC is much more likely to overestimate the number of classes.

Whereas BIC and CAIC perform less well than AIC and AIC3, there is a substantial difference among the two versions of the measures. With BIC(N) and CAIC(N) it is much more likely to underestimate the number of mixture components than with BIC(K) or CAIC(K). This is true for all conditions, except for the very low separation condition ($n_k = 5$).

5.5.3 Overall results

The main goal of our simulation study was to determine which of the investigated model selection measures is preferable for deciding simultaneously about the number of lower and higher-level mixture components in multilevel LC models. Table 5.7 presents the aggregated results for each level of the five design factors. It reports the number of simulation replications in which the number of lower and higher-level mixture components were correctly estimated ($\hat{L} = L$ and $\hat{H} = H$) and incorrectly estimated, either at the lower, higher, or both levels. Note that we also present the results obtained when using BIC and CAIC with sample size N for the lower-level analysis and K for the higher-level analysis.

Comparison of the results for the investigated fit measures shows that overall AIC3 performs best, which confirms what has been found earlier for standard and multilevel LC models (Dias, 2004; Lukočienė and Vermunt, 2009). AIC performs worse than AIC3 except in the case when the higher-level separation is weak ($n_k = 5, 10$). BIC(K, K) and BIC(K, N)

perform similarly well and better than BIC(N, N), and the same applies to the three versions of CAIC. However, BIC performs better than CAIC with the same sample size definitions.

5.6 An empirical example

We will illustrate the three-step model selection procedure with an analysis of one of the yearly surveys among university graduates by the consortium AlmaLaurea. In the current application, we use the questionnaire items on job satisfaction answered by the summer 2004 graduates of the University of Florence (AlmaLaurea, 2006). Information is available for 826 graduates having a job at the moment of interview and belonging to 23 study programs, where the smallest number of graduates per program is 8 and the largest is 155. The 12 dichotomous questionnaire items of interest measure the following aspects of the satisfaction with the current job: stability, coherence with the study, competence/professionalism, prestige, cultural interests, social utility, independence, involvement in the working activity and in the decisional processes, schedule flexibility, salary, and career, as well as the global satisfaction. The aim of the multilevel LC analysis is to cluster graduates into classes differing in their responses to the satisfaction items, as well as to cluster programs based on the distribution of graduates across the graduate-level satisfaction classes.

Table 5.8 summarizes the results of applying our three-step model fitting procedure. In step 1 (ignoring the hierarchical data structure),

Table 5.8: Summary of selected models in step 1, 2, and 3 with the data on the summer 2004 graduated at the University of Florence.

Information criterion	L, step 1	H, step 2	L, step 3
BIC(K,K)	5	2	8
BIC(N,N)	4	2	4
BIC(K,N)	4	2	4
CAIC(K,K)	5	2	6
CAIC(N,N)	4	2	4
CAIC(K,N)	4	2	4
AIC3	8	2	8

BIC(N) and CAIC(N) select the model with 4 lower-level classes, BIC(K) and CAIC(K) the model with 5 classes, AIC3 the model with 8 classes, and AIC the model with 9 classes. For step 2, we estimated multilevel LC models with $L = 4$, $L = 5$, and $L = 8$ (we did not proceed with the AIC result $L = 9$). Irrespective of the value of L and the information criterion that is used, a model with 2 classes at the program level should be preferred. In step 3, we estimated models with 2 LCs at the program level and different numbers of LCs at the graduate level. BIC(N), CAIC(N), and AIC3 select the same solution as in step 1, whereas BIC(K) and CAIC(K) select models with a larger number of lower-level classes (8 and 6, respectively). The explanation for the fact that these criteria select a larger number of lower-level classes in step 3 is that the higher-level separation is very good (ranging from .86 in the model with $L = 6$ to .89 in the model with $L = 8$). Note that BIC(K) and AIC3 come up with the same final conclusion, which can be explained by the fact that their penalties are very similar: $\ln 23 = 3.14$, which is very close to 3.

Of course, not only fit indexes are important for model selection, but also the interpretability of the obtained solutions. Because the solution

Table 5.9: Distribution of student-level classes within program-level classes obtained for the $H = 2$ and $L = 4$ model obtained with the data on the summer 2004 graduated at the University of Florence.

	h	
l	1	2
1	0.63	0.33
2	0.17	0.20
3	0.10	0.25
4	0.09	0.22

with 8 latent classes at the lower level is somewhat difficult to interpret, we will describe the solution with 4 lower-level and 2 higher-level classes. Lower-level class 1 contains the graduates who are satisfied with all aspects of the current job and class 4 the ones who are dissatisfied with all job aspects. The other two classes are satisfied with some and dissatisfied with other aspects: Class 2 is dissatisfied with job stability, salary, and career opportunities, and class 3 with coherence with the study and cultural interests.

At the program level there are two classes, where class 1 is the larger of the two [$P(w_k = 1) = 0.81$]. Table 5.9 shows how the two classes differ in terms of their student-level class membership probabilities $P(x_{kj} = l|w_k = h)$. As can be seen, programs belonging to class 1 score much better in terms of the satisfaction of their graduates than programs belonging to class 2. Compared to the latter, the former have a much larger proportion of graduates belonging to the satisfied lower-level LC one, and a much smaller proportion of students belonging to the dissatisfied lower-level LC four. Also the proportions of graduates in the partially dissatisfied classes two and three are slightly smaller.

5.7 Conclusions

The purpose of the current study on multilevel LC models was twofold, namely, evaluating the performance of a new three-step model fitting procedure and investigating the performance of information criteria for simultaneously deciding about the number of lower- and higher-level classes.

As far as the performance of the three-step procedure is concerned, the simulation study did not provide strong evidence that it is an improvement over the two-step procedure used by Vermunt (2003). A possible explanation for this is that our lower-level models were never very weak. It can be expected that the third step will be more important when lower-level classes are badly separated and higher-level classes very well separated. This is what occurred in the application in which the additional third step turned out to matter. What can be said is that the third step will never harm, but that more research is needed to demonstrate under which circumstances it is really needed.

As far as the sample size definition for BIC and CAIC is concerned, our simulation study showed clearly that the number of groups (K) is the only appropriate sample size for deciding on the number of classes at the higher level, which is in agreement with the results reported by Lukočienė and Vermunt (2009). For the decision about the number of lower-level classes, it makes less of a difference which sample size is used, but somewhat surprisingly also here BIC(K) and CAIC(K) perform slightly better than BIC(N) and CAIC(N).

Overall, AIC3 turns out to be the preferred criterion for simultaneously deciding about the number of lower- and higher-level classes. This is in agreement with simulation results for standard LC models (Andrews and Currim, 2003; Dias, 2004; Sarstedt, 2008). The BIC criterion with sample size K is the second best measure, both for the lower and higher level. In situations with very low separation between higher-level classes, the AIC criterion performs best.

Summary

Latent class (LC) analysis has become one of the standard data analysis tools in applied research areas such as social sciences, behavioral sciences, and the biomedical field. This thesis focused on the use of LC models (also referred to as mixture models) as tools for multilevel analysis; that is, for capturing variation between higher-level units in a nested data structure by assuming that these units belong to homogeneous LCs. Two types of LC models were investigated: LC regression models and multilevel LC models. The former are two-level models and the latter three-level models.

Chapters 2 and 3 dealt with LC regression modeling. This tool can be used for defining (two-level) random-effects regression models with a non-parametric or semi-parametric specification of the random effects distribution. The difference between the nonparametric and semi-parametric approach is that in the latter the number of classes is determined using a particular fit measures (e.g. BIC), whereas the former involves increasing the number of classes till the log-likelihood function no longer increases. One of the aims of this thesis was to compare the performance of these two LC-based random-effects approaches with that of traditional parametric approaches, which typically rely on the assumption that random

effects come from a multivariate normal distribution. The LC regression approach has several practical advantages when applied with categorical response variables, but this is, of course, not enough to prefer this particular method.

Chapter 2 studied the sensitivity of two-level logistic regression analysis for misspecification of the random effects distribution. More specifically, it was investigated whether using a nonparametric specification of the random-effects distribution reduces bias and increases efficiency when random effects are not normally distributed. For moderate intraclass correlations, this turned out to be the case as long as the level-1 sample size is not too small. However, when the level-1 sample size is very small (say three), the standard parametric approach outperformed the nonparametric approach, even when the random effects distribution is misspecified. For small intraclass correlations, the two approaches performed equally well.

Chapter 3 investigated the performance of three types of random coefficients logistic regression models; that is, models using parametric, semi-parametric, and nonparametric specifications of the distribution of the random effects. Whereas earlier studies (including Chapter 2) focused on models with a single random effect, here we looked at models with multidimensional random effects (intercepts and slopes). Moreover, also the performance of a semi-parametric approach – using LC regression models where the number of LCs is selected using the BIC – was investigated. One of the main conclusions of Chapter 3 was that the good results obtained with the nonparametric approach in the unidi-

mensional case do not generalize to the multidimensional case. Parametric and semi-parametric approaches are much better in terms of bias and relative efficiency than the nonparametric approach. For the fixed-effects estimation, a parametric approach is the preferred method when the underlying assumption of the parametric model holds. In all other situations, the semi-parametric approach is the best choice.

Chapter 4 and 5 dealt with multilevel LC models. These models are multilevel extensions of the standard LC model itself. The aim of these models is to build a meaningful cluster model for the lower-level units as well as to cluster higher-level units based on the distribution of their members across the lower-level clusters. A complicating issue when applying these models is that they require the simultaneous decision about the number of classes at multiple hierarchical levels. Little is known on how to decide about the number of LCs at each of the two hierarchical levels. Moreover, it is unknown how to deal with the fact that these decisions are dependent of one another.

In Chapter 4 the performance of various model selection methods was investigated in the context of multilevel mixture models. This chapter focused on determining the number of mixture components at the higher-level under the somewhat simplified situation that the number of lower-level classes is known. We considered information criteria BIC, AIC, and AIC3, and CAIC as well as ICOMP and the validation log-likelihood. A specific difficulty that occurs in the application of BIC and CAIC in the context of multilevel models is that they contain the sample size as one of their terms and it is not clear which sample size should be used in their

formula. This could be the number of groups, the number of individuals, or either the number of groups or number of individuals depending on whether one wishes to determine the number of components at the higher or at the lower level. Our simulation study showed that when one wishes to determine the number of mixture components at the higher level, the most appropriate sample size for BIC and CAIC is the number of groups (higher-level units). Moreover, it was found that BIC, CAIC and ICOMP detect very well the true number of mixture components when both the components' separation and the group-level sample size are large enough. AIC performs best with low separation levels and small sizes at the group-level.

Finally, Chapter 5 expanded the study of Chapter 4 by proposing a new three-step model fitting procedure for simultaneously deciding about the number of higher- and lower-level classes, as well as by investigating the performance of information criteria (BIC, AIC, CAIC and AIC3) when also the number of lower-level classes is unknown. The three main conclusions of the simulation study were that 1) the proposed three-step model fitting strategy works rather well, 2) the number of higher-level units is the preferred sample size for BIC and CAIC, both for decisions about higher- and lower-level classes, and 3) AIC3 is the preferred measure for deciding about the number of LCs both at the higher and lower level, except for situations with very badly separated (higher-level) classes, in which case AIC performs best.

Bibliography

Agresti, A., B. Caffo and P. Ohman-Strickland (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis* **47**, 639–653.

Agresti, A., J.G. Booth, J.P. Hobert and B. Caffo (2000). Random-effects modeling of categorical response data. *Sociological Methodology* **30**, 27–80.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 218–234.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

AlmaLaurea (2006). Condizione occupazionale dei laureati pre e post riforma, viii indagine 2005. Technical report. Bologna: Consorzio Interuniversitario AlmaLaurea.

Andrews, R.L., A. Ansari and I.S. Currim (2002*a*). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research* **39**, 479–487.

Andrews, R.L., A. Ansari and I.S. Currim (2002*b*). Hierarchical bayes versus finite mixture conjoint analysis models: a comparison of fit, prediction, and partworth recovery. *Journal of Marketing Research* **39**, 87–98.

Andrews, R.L. and I.S. Currim (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research* **40**, 235–243.

Asparouhov, T. and B.O. Muthén (2008). Multilevel mixture models. In: *Advances in latent variable mixture models* (G.R. Hancock and K.M. Samuelsen, Eds.). pp. 27–75. Charlotte, NC: Information Age Publishing, Inc.

Bartholomew, D.J. and M. Knott (1999). *Latent variable models and factor analysis*. London: Arnold.

Bassi, F. (2009). Latent class models for marketing strategies: an application to the italian pharmaceutical market. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* **5**, 40–45.

Bijmolt, T.H.A., L.J. Paas and J.K. Vermunt (2004). Country and consumer segmentation: multilevel latent class analysis of financial product ownership. *International Journal of Research in Marketing* **21**, 323–340.

Bock, R.D. and M. Aitkin (1981). Marginal maximum likelihood estimation of item parameters. *Psychometrika* **46**, 443–459.

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping and others*. London: Chapman & Hall.

Bouwmeester, S., J.K. Vermunt and K. Sijtsma (2007). Development and individual differences in transitive reasoning: a fuzzy trace theory approach. *Developmental Review* **27**, 41–74.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): the general theory and its analytical extensions. *Psychometrika* **52**, 345–370.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. In: *Information and Classification* (O. Opitz, B. Lausen and R. Klar, Eds.). Springer, Heidelberg. pp. 218–234.

Breslow, N.E. and D.G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Bryk, A.S. and S.W. Raudenbush (1992). *Hierarchical linear models: application and data analysis methods*. Newbury Park, CA: Sage Publications.

Busing, F. (1993). Distribution characteristics of variance estimates in two-level models. Unpublished manuscript. Department of Psychometrics and Research Methodology, Leiden University.

Cavrinia, G., G. Galimbertia and G. Soffritti (2009). Evaluating patient satisfaction through latent class factor analysis. *Health & Place* **15**, 210–218.

Clogg, C.C. and L.A. Goodman (1984). Latent structure analysis of a set of multidimentional contingency tables. *Journal of the Amerincan Statistical Association* **79**, 762–771.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood estimation from incomplete data via the em algorithm. *Journal of the Royal Statistical Society* **39**(1), 1–38.

Di, C. and K. Bandeen-Roche (2008). Multilevel latent class models with dirichlet mixing distribution. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 174. http://www.bepress.com/jhubiostat/paper174.

Dias, J.G. (2004). Finite mixture models: review, applications and computer intensive methods. PhD thesis. Doctoral dissertation, SOM Research School, University of Groningen, Netherlands.

Dias, J.G. (2006). Model selection for the binary latent class model: A monte carlo simulation. In: *Data science and classification* (V. Batagelj, H.-H. Bock, A. Ferligoj and A. ˇ Ziberna, Eds.). Springer, Berlin. pp. 91–99.

Fox, J.P. and C.A.W. Glas (2001). Bayesian estimation of a multilevel irt model using gibbs sampling. *Psychometrika* **66**, 269–286.

Galindo-Garre, F., J.K. Vermunt and W. Bergsma (2004). Bayesian posterior estimation of logit parameters with small samples. *Sociological Methods and Research* **39**(33), 88–117.

Goldstein, H. (1995). *Multilevel statistical models*. New York: Halsted Press.

Hagenaars, J.A. and A.L. McCutcheon (2002). *Applied latent class analysis models*. Cambridge University Press.

Hartzel, J., A. Agresti and B. Caffo (2001). Multinomial logit random effects models. *Statistical Modelling* **1**, 81–102.

Heagerty, P.J. and B.F. Kurland (2001). Misspecified maximum likelihood estimates and generalized liner mixed models. *Biometrica* **88**, 973–985.

Heckman, J.J. and B. Singer (1982). Population heterogeneity in demographic models. In: *Multidimensional mathematical demography* (K.C. Land and A. Rogers, Eds.). New York: Academic Press. pp. 567–599.

Heckman, J.J. and B. Singer (1984). A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrica* **52**, 271–320.

Hedeker, D. and R.D. Gibbons (1996). Mixor: A computer program for mixed effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* **49**, 157–176.

Hox, J. (2002). *Multilevel analysis: techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J.J. (1994). *Applied multilevel analysis*. Amsterdam: TT.

Hox, J.J. and C.J.M. Maas (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling* **8**, 157–174.

Huq, N.M. and J. Cleland (1990). Bangladesh fertility survey 1989 (main report). Technical report. Dhaka: National Institute of Population Research and Training.

Kragelj, B. and E. Schlutter (2007). "digital divide" reconsidered: a country- and individual-level typology of digital inequality in 26 european countries. V: Quantitative methods in the social sciences. Programme and abstracts. Prague: European Science Foundation.

Kreft, I.G.G. and J. de Leeuw (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage Publications.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixture distribution. *Journal of the American Statistical Association* **73**, 805–811.

Lee, Y. and J.A. Nelder (2004). Conditional and marginal models: another viewn. *Statistical Science* **19**, 219–228.

Leisch, F. (2004). Flexmix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software*.

Leroux, B.G. (1992*a*). Consistent estimation of a mixing distribution. *Annals of Statistics* **20**, 1350–1360.

Leroux, B.G. (1992*b*). Maximum likelihood estimation for hidden markov models. *Stochastic Process. Appl.* **40**, 127–143.

Leroux, B.G. and M.L. Puterman (1992). Maximum-penalized-likelihood estimation for independent and markov-dependent mixture models. *Biometrics* **48**, 545–558.

Lesaffre, E. and B. Spiessens (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example.. *Applied Statistics* **50**, 325–335.

Lin, T.S. and C.M. Dayton (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics* **22**, 249–264.

Lindsay, B.G. (1983). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics* **11**, 86–94.

Lindsay, B.G. (1995). Mixture models: theory, geometry and applications. In: *NSF-CBMS Regional Conference Series in Probability and Statistics*. Vol. 5. Hayward, CA: Institute of Mathematical statistics.

Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 817–827.

Longford, N.T. (1995). *Random coefficient models*. Oxford: Clarendon.

Lukočienė, O. and J.K. Vermunt (2008). A comparison of multilevel logistic regression models with parametric and nonparametric random intercepts. Manuscript submitted for publication.

Lukočienė, O. and J.K. Vermunt (2009). Determining the number of components in mixture models for hierarchical data. In: *Advances in data analysis, data handling and business intelligence* (A. Fink, L. Berthold, W. Seidel and A. Ultsch., Eds.). pp. 241–249. Berlin-Heidelberg: Springer.

Maas, C.J.M. and J.J. Hox (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis* **46**, 427–440.

Magder, L.S. and S.L. Zeger (1996). A smooth nonparametric estimate of mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association* **91**, 1141–1151.

Magidson, J. and J.K. Vermunt (2004). Latent class models. In: *The Sage Handbook of Quantitative Methodology for the Social Sciences* (D. Kaplan, Ed.). pp. 175–198. Thousand Oakes: Sage Publications.

Magidson, J. and J.K. Vermunt (2007). Use of a random intercept in latent class regression models to remove response. In: *Bulletin of the International Statistical Institute, 56th Session.* number 1604. pp. 1–4.

McLachlan, G. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**, 318–324.

McLachlan, G. and D. Peel (2000). *Finite mixture models*. New York: J. Wiley and Sons. Inc.

Múthen, L.K. and B.O. Múthen (1998). *Mplus user's guide*. Los Angeles, CA: Muthen & Muthen.

Múthen, L.K. and B.O. Múthen (2006). *Mplus user's guide*. fourth edition ed.. Los Angeles, CA: Muthen & Muthen.

Neuhaus, J.M., W.W. Hauck and J.D. Kalbfleisch (1992). The effects of mixture distribution misspecification when fitting mixed effects logistic models. *Biometrika* **79**, 755–762.

Nylund, K.L., B.O. Múthen and T. Asparouhov (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal* **14**, 535–569.

Palardy, G. and J.K. Vermunt (in press). Multilevel growth mixture models for classifying group-level observations. *Journal of Educational and Behavioral Statistics*.

Pan, J.X. and R. Thompson (2003). Gauss-hermite quadrature approximation for estimationin in generalised linear mixed models. *Computational Statistics* **18**, 57–78.

Pirani, E., S.S. DíAndrea and J.K. Vermunt (2009). Poverty and social exclusion in europe: differences and similarities across regions. Paper presented at the XXVI IUSSP Conference, Marrakech.

Rabe-Hesketh, S., A. Pickles and A. Skrondal (2001). Gllamm: A general class of multilevel models and a stata program. *Multilevel Modelling Newsletter* **13**, 17–23.

Rabe-Hesketh, S., A. Pickles and A. Skrondal (2003). Correcting for covariate measurement error in logistic regresson using nonparametric maximum likelihood estimation. *Statistical Modelling* **3**, 215–232.

Rabe-Hesketh, S., A. Pickles and A. Skrondal (2004). Generalized multilevel structural equation modeling. *Psychometrica* **69**, 167–190.

Rabe-Hesketh, S., A. Skrondal and A. Pickles (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* **128**, 301–323.

Rindskopf, D. (2006). Heavy alcohol use in the "fighting back" survey sample: separating individual and community level influences using multilevel latent class analysis. *Journal of Drug Issues* **36**, 441–462.

Rodriguez, G. and N. Goldman (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society Series A* **158**, 73–89.

Rodriguez, G. and N. Goldman (2001). Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society Series A* **164**, 339–355.

Sarstedt, M. (2008). Market segmentation with mixture regression models: understanding measures that guide model selection. *Journal of Targeting, Measurement and Analysis for Marketing* **16**(3), 228–246.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**(2), 461–464.

Singer, J.D. (1998). Using sas proc mixed to fit multilevel models, hierarchical models, and individual growth models8).. *Journal of Educational and Behavioral Statistics* **24**, 323–355.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized latent variables modeling: multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hal/CRC.

Smyth, D. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing* **9**, 63–72.

Snijders, T.A.B. and R.J. Bosker (1999). *Multilevel analysis*. London: Sage Publications.

Stiratelli, R., N. Laird and J.H. Ware (1984). Random-effects models for serial observations with binary responses. *Biometrics* **40**, 961–971.

Verbeke, G. and G. Molenberghs (2000). *Linear mixed models for longitudinal data*. Springer, Berlin.

Vermunt, J.K. (1997). *Log-linear models for event histories*. . Thousand Oakes, CA: Sage Publications.

Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology* **33**, 213–239.

Vermunt, J.K. (2004). An em algorithm for the estimation of paramet-

ric and nonparametric hierarchical nonlinear models. *Statistical Neerlandica* **58**, 220–233.

Vermunt, J.K. (2005). Mixed-effects logistic regression models for indirectly observed outcome variables. *Multilevel Behavioral Research* **40**, 281–301.

Vermunt, J.K. (2007). A hierarchical mixture model for clustering three-way data sets. *Computational Statistics and Data Analysis* **51**, 5368–5376.

Vermunt, J.K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research* **17**, 33–51.

Vermunt, J.K. and J. Magidson (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced.*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K. and J. Magidson (2008). *LG-syntax user's guide: manual for Latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K. and L. Van Dijk (2001). A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modelling Newsletter* **13**, 6–13.

Wedel, M. (2001). Glimmix: Software for estimating mixtures and mixtures of generalised linear models. *Journal of Classification* **18**, 129–135.

Wedel, M. and W. A. Kamakura. (1998). *Market segmentation: concepts and methodological foundations*. Boston: Kluwer Academic Publishers.

Wedel, M. and W.S. DeSarbo (1994). A review of recent developments in latent class regression models. In: *Advanced Methods of Marketing Research* (R.P. Bagozzi, Ed.). Cambridge: Blackwell Publishers. pp. 352–388.

Wolfinger, R. and M. O'Connell (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–243.

Wong, G.Y. and W.M. Mason (1985). Hierarchical logistic models for multilevel analysis. *Journal of the American Statistical Association* **80**, 513–524.

Wood, A. and J. Hinde (1987). Binomial variance component models with a non-parametric assumption concerning random effects. In: *Longitudinal Data Analysis: Surrey Conference on Sociological Theory and Method 4*. Avebury, Aldershot: Hants. pp. 110–128.

Zeger, S.L., K. Liang and P.S. Albert (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.